University of Sheffield

Metadiscourse Tagging in Academic Lectures

By

Ghada ALharbi

A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy

in the

Speech and Hearing Research Group
Department of Computer Science

August 2016

Declaration

I hereby declare that this thesis is of my own composition, and that it contains no material previously submitted for the award of any other degree. The work reported in this thesis has been executed by myself, except where due acknowledgement is made in the text.

Ghada AlHarbi

August 2016

"Faith is a knowledge within the heart, beyond the reach of proof."

Khalil Gibran

Abstract

This thesis presents a study into the nature and structure of academic lectures, with a special focus on metadiscourse phenomena. Metadiscourse refers to a set of linguistics expressions that signal specific discourse functions such as the Introduction: "*Today we will talk about...*" and Emphasising: "*This is an important point*". These functions are important because they are part of lecturers' strategies in understanding of what happens in a lecture. The knowledge of their presence and identity could serve as initial steps toward downstream applications that will require functional analysis of lecture content such as a browser for lectures archives, summarisation, or an automatic minute-taker for lectures. One challenging aspect for metadiscourse detection and classification is that the set of expressions are semi-fixed, meaning that different phrases can indicate the same function.

To that end a four-stage approach is developed to study metadiscourse in academic lectures. Firstly, a corpus of metadiscourse for academic lectures from Physics and Economics courses is built by adapting an existing scheme that describes functional-oriented metadiscourse categories. Second, because producing reference transcripts is a time-consuming task and prone to some errors due to the manual efforts required, an automatic speech recognition (ASR) system is built specifically to produce transcripts of lectures. Since the reference transcripts lack time-stamp information, an alignment system is applied to the reference to be able to evaluate the ASR system. Then, a model is developed using Support Vector Machines (SVMs) to classify metadiscourse tags using both textual and acoustical features. The results show that n-grams are the most inductive features for the task; however, due to data sparsity the model does not generalise for unseen n-grams. This limits its ability to solve the variation issue in metadiscourse expressions. Continuous Bag-of-Words (CBOW) provide a promising solution as this can capture both the syntactic and semantic similarities between words and thus is able to solve the generalisation issue. However, CBOW ignores the word order completely, something which is very important to be retained when classifying metadiscourse tags.

The final stage aims to address the issue of sequence modelling by developing a joint CBOW and Convolutional Neural Network (CNN) model. CNNs can work with continuous features such as word embedding in an elegant and robust fashion by producing a fixed-size feature vector that is able to identify indicative local information for the tagging task. The results show that metadiscourse tagging using CNNs outperforms the SVMs model significantly even on ASR outputs, owing to its ability to predict a sequence of words that is more representative for the task regardless of its position in the sentence. In addition, the inclusion of other features such as part-of-speech (POS) tags and prosodic cues improved the

results further. These findings are consistent in both disciplines. The final contribution in this thesis is to investigate the suitability of using metadiscourse tags as discourse features in the lecture structure segmentation model, despite the fact that the task is approached as a classification model and most of the state-of-art models are unsupervised. In general, the obtained results show remarkable improvements over the state-of-the-art models in both disciplines.

Acknowledgements

In the name of Allah, the Most Gracious and the Most Merciful. Thanks to Allah who is the source of all the knowledge in this world, and imparts as much as He wishes to anyone He finds suitable. I would like to thank my supervisor, Professor Thomas Hain, who has supported me throughout my work on this thesis with his great knowledge, advice and guidance. I have been extremely lucky to be part of the Machine Intelligence for Natural Interface (MINI) research group that offer a great working environment and infrastructure. Special thanks to my second supervisor, Professor Robert Gaizauskas for his very useful suggestions, support and for being so kind to me. I am indebted to Raymond W. M. Ng and Oscar Saz Torralba who helped me throughout my work and gave me very useful suggestions regarding it. I am also grateful to Yulan Liu for her friendship, support and encouraging words during my PhD study.

I am thankful to my parents for their support, prayers, love and care throughout my life and they have played a vital role in achieving this milestone: to my brothers and sisters and to my special friend Nouf Alharbi who has always been a wonderful being to me with her limitless support during my study. Also, special thanks to Roaa AlDossary and Rania AlGhamdi who extended their support especially during my PhD studies. Finally, I am thankful to the Royal Embassy of Saudi Arabia Cultural Bureau in London, for funding this work.

Ghada Fahad Alharbi, 31 August, 2016

Contents

List of Figures

List of Tables

Dedicated to my parents...

Chapter 1

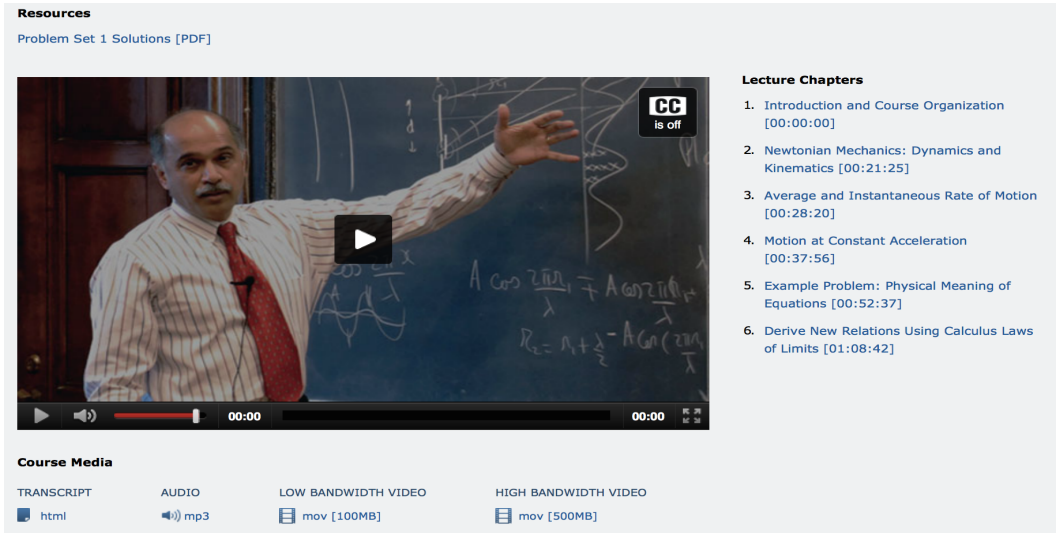
Introduction

Knowledge about discourse structure is one of the most important topics in both natural language processing and spoken language understanding, and has been extensively studied over the last decade; however, it is still in its early stages, especially for realistic and complex scenarios such as academic lectures. The challenge mainly lies in the large variations in speaker style, lecture topics, and noise in this type of data. The task of discourse analysis in speech comprises many aspects of interaction interpretation, including meanings, rhetoric and actions. This thesis studies the discourse rhetoric interpretation problem, specifically focusing on certain expressions that lecturers use to explicitly mark the rhetorical functions of utterances with respect to the global discourse structure. These expressions are often referred to as metadiscourse. Work with a thematic discourse segmentation application for academic lectures is evaluated in this study.

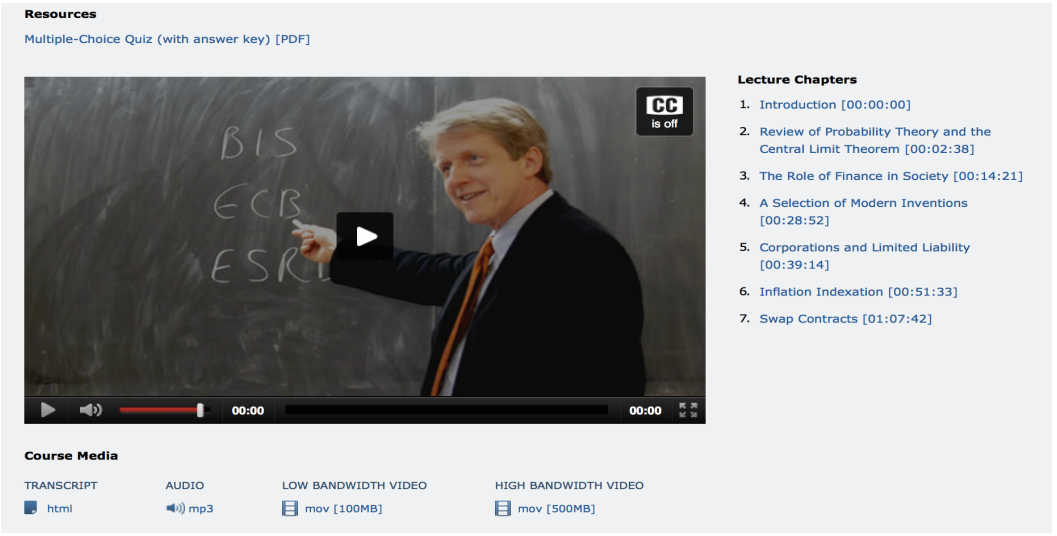
In this chapter, the motivations for the work and its potential applications are presented, along with an outline of the research focus. The main contributions of the work are further defined, and finally, the structure of the thesis is presented.

1.1 Motivation

Finding a formal description of the structures of academic lectures from different disciplines can be of practical use for various educational applications, such as summarisation and interactive webcasting. The specific motivation for finding such descriptions in this work is to improve the thematic segmentation model of online lecture courses. The thematic segmentation application aims to provide an outline of a lecture's content, based on its segment themes, which are either topical or functional. A concrete example of lecture structure is presented in Figure 1.1, where the outlines of both the Physics and Economics lectures are



(A)



(B)

FIGURE 1.1: Examples of the Interface used for browsing (a) physics and (b) economics lectures in OYC, content provided by Ramamurti (2006) and Shiller (2011).

organised as chapters. Some chapters represent an introduction of the lecture, such as the first chapters in Figures 1.1 (A) and (B); some provide a particular example, as in the fifth chapter of Figure 1.1 (A); others review a specific point, as in the second chapter of Figure 1.1 (B); and others simply introduce new topics, as in the rest of the chapters in both (A) and (B). These two examples indicate that lecture segments do not just represents topics, they also demonstrate rhetorical functions that reflect the pedagogical nature of academic lectures. This formulation situates the approach of this thesis in the area of discourse structuring and segmentation.

Segmentation concerns splitting the content of a document into cohesive segments that represent a meaningful structure – in this case the themes and functions within academic lectures. Many statistical algorithms have been proposed to address this issue by tracking the dramatic changes in vocabulary usage within the document (Galley et al., 2003, Hearst,

1997, Mohri et al., 2010). Such changes in vocabulary are commonly known as lexical cohesion (Morris and Hirst, 1991). Earlier studies have approached the segmentation task based on certain discourse indicators known as cue phrases, such as 'now', 'so', or 'well' (Grosz and Sidner, 1986, Hirschberg and Litman, 1993b). Other studies have addressed both lexical cohesion and discourse cues in the segmentation algorithm (Eisenstein and Barzilay, 2008, Galley et al., 2003), which show great improvements over previous models, particularly for written discourse, though less so for spoken discourse. One possible reason for low performance in the analysis of lectures is that most of the state-of-the-art segmentation models do not consider the rhetorical functions of lecture discourse in the segmentation algorithms. It is nonetheless a big challenge to represent lecture content efficiently based on both functions and topics, as it requires a level of understanding of the lecture content to locate these functional regions in the discourse. This might involve applying a suitable process to the transcripts of audio/visual media, to locate regions indicating introductions, examples or reviews, through rhetorical functions in the transcripts, without fully understanding the text.

Researchers in the field of English language learning show that lecturers use recognised strategies in the form of linguistic expressions to plan and organise the content of their lectures. The purpose of these strategies is to explicitly direct the students in the realm of the communication event. Such expressions are commonly referred to as metadiscourse (also known as *metalanguage* or *signposting language*), and defined by Crismore et al. (1993) as "linguistic material in texts, written or spoken, which does not add anything to the propositional content but that is intended to help the listener or reader organise, interpret and evaluate the information given". Examples of metadiscourse functions include highlighting important concepts in the lecture ('This is very important to understand...'); demonstrating something with examples ('Let's take an example of...'); or reviewing some ideas from previous lectures ('Last lecture I introduced the concept of...'). A core characteristic of metadiscourse is that it does not contribute to the topic of the lectures. Another interesting property is that it occurs in both written and spoken discourse. By identifying and tagging metadiscourse according to its function, similar content to that given in Figure 1.1 can be generated automatically, to enhance the navigation experience of the student, and ultimately be a part of a learning framework for delivering lecture content.

1.2 Research Focus

There are two main questions that arise when conducting the metadiscourse tagging task for academic lectures that could benefit several downstream applications, including thematic discourse segmentation. Firstly, **how should one identify and classify instances of metadiscouse in academic lectures**? To answer this question, this study will investigate

to what extent hired annotators are able to understand the task, in order to build a corpus of metadiscouse in academic lectures from different disciplines. Secondly, **to what extent can one develop a robust metadiscourse tagging model that is capable of dealing with the issue of expressions variation**? More specifically, this thesis aims to obtain insight into the nature of the metadiscourse phenomenon itself, by understanding what features or feature combinations are most representative of it in lectures, and how different models perform when conducting the task. Additionally, this question addresses issues of performance when Automatic Speech Recognition (ASR) outputs are used instead of manual transcripts, where the number of misclassification instances can be severe. Overall, this thesis focuses on these two areas related to corpus building, and the modelling paradigm for the resulting datasets. The problem is important, as it is at the heart of many discourse-based applications.

To answer these two questions, a four-stage approach is developed, involving corpus building, feature representation, and a classification model that is able to capture metadiscourse expressions. First, an annotation scheme is defined based on existing theoretical perspectives, which provides reliable coding strategies for metadiscourse in academic lectures across different disciplines. The scheme chosen consists of 19 distinctive tags that can be mapped under 4 general ones. In stage 2, an ASR system for academic lectures is built, to automate the process of producing lecture transcriptions. In addition, a strategy is proposed to transfer the gold standard tags from the reference transcription to the ASR outputs, in order to train the model on them. In stage 3, using both textual and acoustic-based features, a classification model based on a Support Vector Machine (SVM) is developed to automate the process of annotating the metadiscourse corpus. In stage 4, using continuous representations of the previously defined features, another classification model is developed based on a Convolutional Neural Network (CNN), to improve the metadiscourse classification performance over the SVM model. The main intuition for using the CNN is that its provides an elegant and robust way of producing a fixed-size feature vector that is able to identify indicative local information for the tagging task. The final contribution with respect to this thesis is to investigate the suitability of using metadiscourse tags at its two levels of granularity (general and specific) as discourse features in the thematic segmentation model of academic lectures for two different disciplines.

1.3 Thesis Contributions

This section lists the contributions made by this thesis, describing the motivation and need for the work carried out:

Contribution 1: Metadiscourse Tagging Approach for Academic Lectures

Motivation. Enriching academic lecture transcripts with metadiscourse tags can be of practical use for downstream applications that require functional analysis, such as a browser for lectures archives, summarisation, or automatic minute-taking for lectures. The importance of metadiscourse comes from the fact that it is part of the lecturers' strategies to assist understanding of what happens in a lecture, and can serve as a guide within the communication event. However, one challenging aspect of metadiscourse is that the set of expressions are semi-fixed, meaning different phrases can indicate the same function. This is because lectures exhibit many variations and challenges due to speaker style, lecture topics and discipline knowledge. Unfortunately, metadiscourse in academic lectures and automating the process of identification and classification has not been widely studied before.

Contribution. This thesis develops a four-stage approach to enrich lecture content with metadiscourse tags, using a robust tagging model that utilises both discrete and continuous features representation, and different classification algorithms. This tagging model was trained on a corpus that was annotated specifically for academic lectures from two different disciplines, with the purpose of showing the effects of the domain on model performance. The work developed can help to reduce the intensive labour required for preparing and organising online lecture materials. Additionally, it provides a complete system in which the input is the recordings of the lectures and the output is automatically generated transcripts from the ASR model, enriched with tags. Evaluation of the thematic discourse segmentation shows that the developed framework attained a significant improvement over state-of-the-art discourse segmentation models. The analysis carried out was based on different test cases, to show which metadiscourse tags most improved the model performance.

Contribution 2: A Corpus of Metadiscourse for Academic Lectures

Motivation. Recently, corpora of metadiscourse have been introduced into multiple natural language applications, such as for the purpose of activity-based summarisation (Niekrasz, 2012), and for building presentation skills tools (Correia et al., 2014a,b). They share the same principles in building a corpus: that is, at first finding a formal definition of the annotation scheme designed for the target task and speech genre (*i.e.:* lectures or meetings), and then collecting and preparing the datasets for the annotation task. They must also decide on the target annotators – either expert or non-expert. Finally, the annotation experiment is conducted with the help of tools and instructions to facilitate the work for the annotators, and to ensure high quality annotation. However, these corpora, as mentioned, were built to serve a specific task for a specific speech genre. As a result, there is no metadiscourse corpus

for academic lectures that could serve the objectives of this thesis' metadiscourse tagging approach for online lecture courses.

Contribution. The first stage of the metadiscourse tagging approach is to build a corpus of metadiscourse in academic lectures. This is accomplished by using an annotation scheme designed to express function-oriented categories at sentence-level, such as the one proposed by Ädel (2010). The scheme allows the grouping of the metadiscourse categories at two level of granularity: generic (contains 4 categories) and specific (contains 19 categories). Each one of the four generic groups can serve as general discourse functions, and these are: Metalinguistic Comments, Discourse Organisation, Speech Acts, and Reference to Audience. Providing both levels of metadiscourse allows one to report on and compare the performance of the tagging model, using both generic and specific metadiscourse categories. Experiments with the selected OpenCourseWare (OCW) lectures datasets show that expert annotators are able to identify occurrences of multiple categories of metadiscourse, and hence confirm a reliable coding of metadiscourse in academic lectures using the adapted annotation scheme.

Contribution 3: Automatic Transcriptions of Academic Lectures

Motivation. The annotation task uses reference transcription in generating metadiscourse tags for each sentence. Producing such manual transcripts is a time-consuming task, and often prone to errors, such as when the transcriber is not aware of the technical words used, and understands or uses other terminology instead (Hazen, 2006). To simplify it, previous research has usually built ASR systems for lectures, such as the one proposed by Glass et al. (2007). Two main components in any ASR system are the acoustic model (AM) and language model (LM). Since these models are usually trained on datasets that differ from the test set, this can cause a mismatch problem that seriously degrades system performance. There are various adaptation techniques that can be applied to the AM or LM, or both, to solve the mismatch problem. However, the choice between them depends largely on the task. For instance, most of the proposed LM adaptation techniques for lectures have relied on the existence of some in-domain materials, such as slides that are associated with the lectures, or on the use of techniques that are able to recognise the slide text from the video, such as optical character recognition (OCR), which often suffers from errors in the recognition of texts.

Contribution. The second stage develops an ASR system for OCW academic lectures with the aim of producing high-quality automatic transcriptions. The system focuses on an LM adaptation that uses a linear interpolation technique by combining both in-domain and out-of-domain materials. The in-domain resources are derived from a set of academic lectures from a wide range of disciplines, selected specifically to be similar to the target lectures in

order to reduce the effects of the mismatch problem. The out-of-domain materials are a large collection of written resources collected from the web. This approach performed remarkably better than other adaptation techniques. It is also considered simple, fast, scalable and competitive. To improve the results further, the AM is trained using in-domain datasets. The model is based on a Deep Neural Network (DNN) combined with a Hidden Markov Model. Finally, a lightly supervised alignment model is applied, which has the benefit of both correcting some errors in the reference transcriptions, and providing time-information in order to evaluate the ASR output.

Contribution 4: Exploring Features with SVMs for Metadiscourse Tagging

Motivation. Many sentence models have been introduced to classify sentences/utterances in both NLP and spoken language understanding tasks, including metadiscourse tagging. They all share the same principles of defining a set of representative features for the task and a classification model. Various feature types have been defined to serve this purpose. For example, n-grams features have proved to be effective in the closely related task of metadiscourse tagging in TED Talks (Correia et al., 2014a). In addition, there are other feature types that are based on acoustic factors, such as prosodic cues, which have not been used before for metadiscourse tagging, but usually improve the classification performance in a number of sentence-level classification tasks, such as dialogue acts tagging (Shriberg et al., 1998, Stolcke et al., 2000, Venkataraman et al., 2002). It will be interesting to investigate whether for metadiscourse tagging the inclusion of prosodic features is complementary to textual features, or may improve the model performance, or have no effect at all.

Contribution. The third stage of the metadiscourse tagging approach presents the baseline tagging model using two kinds of feature sets: text-based and acoustic-based. Finding the best combination of features set for the task is one of the primary focuses in this work, due to the significant effect this choice has on the model performance. In order to combine high-dimensional (*e.g.,* words n-grams) with low-dimensional features (*e.g.,* prosodic cues), a support vector machine (SVM) is used, which allows easy integration of both modalities. This is because SVMs can learn independently of the dimensionality of the feature space, by measuring the complexity of hypotheses based on the margin by which they separate the data points, and do not depend on the number of features (Joachims, 1998). To prove the robustness of the model developed, a number of test cases are set out for that purpose, including generic and specific metadiscourse tags, and the ASR output. Evaluation of these test cases shows the ability of the model to classify metadiscourse categories. The main drawback of the model with such features is the sparsity problem, as the task of metadiscourse tagging relies on the existence of certain expressions in the sentence. A large labelled dataset is required to cover all the different variants of metadiscourse expressions.

Contribution 5: Improving Metadiscourse Tagging with CNNs

Motivation. Previous approaches that used a combination of hand-engineered features with SVMs suffered from sparsity problems, and the current model will not be able to generalise well for unseen *n*-grams. This limits its ability to solve the variants issue in metadiscourse expressions, thus limiting the effectiveness of this method. This can be solved by using Continuous Bag-of-Words (CBOW), as it can capture both the syntactic and semantic similarities between words, and thus solve the generalisation issue. A downside of CBOW is that it ignores the word order completely, something which is very important to retain when classifying metadiscourse tags.

Contribution. To solve this issue, the final stage uses a Convolutional Neural Network (CNN) for the metadiscourse tagging task. CNNs are designed to identify the most inductive features of an order of items locally, regardless of their positions in the sentence, for the classification task at hand. A key aspect of using CNN models is to represent the features using dense, low-dimensional vectors, instead of sparse, high-dimensional vectors. Another interesting property of CNNs is the ability to capture non-linear interactions between feature vectors. This model is evaluated using the same metric and test setup as in the previous model. Again, to prove the robustness of the developed model, a number of test cases were set, including testing the model at two functional levels of metadiscourse tags (generic and specific). Experiments using the pre-trained word embedding vectors from *word2vec* and *GloVe* show remarkable performance improvements over the traditional approach using SVM models. Moreover, the inclusion of prosodic features along with Part-of-Speech tags improve the model further.

1.4 Thesis Overview

The remainder of this thesis is presented in Chapters 2 to 8. The content of these chapters is summarised below.

- **Chapter 2: Metadiscourse Tagging Approach for Academic Lectures**
  This chapter presents the approach for metadiscourse tagging over four stages, each representing a different task component. A breakdown of this approach with more detail is depicted in the subsequent chapters. This chapter justifies contribution 1.

- **Chapter 3: Annotating Metadiscourse in Academic Lectures**
  This chapter presents the first stage of the approach developed, which involved building a corpus of metadiscourse for academic lectures by adopting the scheme from Ädel (2010). It also presents the effects of discipline knowledge in two distinct types of

lecture course: Physics and Economics. The contents of this chapter are based on the papers Alharbi and Hain (2016) and Alharbi et al. (2015), and justify contribution 2.

- **Chapter 4: Automatic Transcriptions of Academic Lectures**
  This chapter introduces an automatic speech recognition (ASR) system for academic lectures. It is focused on language model (LM) adaptation, in particular the linear interpolation of in-domain and out-of-domain resources to improve the ASR performance. Lightly supervised alignment techniques are also applied to the reference transcripts to enable evaluation and scoring against ASR outputs. The parts of the ASR model used here are based on Alharbi and Hain (2012) and W. M. Ng et al. (2015), and the contents of this chapter justify contribution 3.

- **Chapter 5: Exploring Features for Metadiscourse tagging with SVMs**
  This chapter presents the first automatic system for classifying metadiscourse instances in lectures at two levels of tags (generic and specific) using a combination of textual and acoustic features. It also demonstrates experiments on the effects of the domain on the classification model. Experiment results investigating the effects of ASR outputs on the classification model performance are also presented. The contents of this chapter justify contribution 4.

- **Chapter 6: Improving Metadiscourse Tagging with CNNs**
  This chapter presents an improvement over the previous classification model for metadiscourse tagging. It is based on the use of both continuous representation of features and CNNs for the tagging task. The contents of this chapter justify contribution 5.

- **Chapter 7: Exploiting Metadiscourse Tags for Discourse Segmentation**
  This chapter demonstrates the results of using metadiscourse tags in a thematic discourse segmentation task for academic lectures. It also compares this approach against the state-of-the-art-model based on lexical cohesion criteria.

- **Chapter 8: Conclusion**
  The last chapter concludes this research by providing a summary, recommendations, and suggestions for the direction of future work.

1.5 Published Work

1. Automatic Transcription of Academic Lectures from Diverse Disciplines. *Ghada Al-Harbi and Thomas Hain, In the Spoken Language Technology Workshop (SLT)*, 2012

2. The USFD spoken language translation system for IWSLT 2014. *Raymond W. M. Ng, Mortaza Doulaty, Rama Doddipatla, Wilker Aziz, Kashif Shah, Oscar Saz, Madina*

Hasan, Ghada AlHarbi, Lucia Specia, and Thomas Hain. In the 11th International Workshop on Spoken Language Translation (SLT), 2015

3. Using Topic Segmentation Models for the Automatic Organisation of MOOCs Resources, *Ghada AlHarbi and Thomas Hain*, *In the 8th International Conference on Education Data Mining (EDM2015)*, 2015

4. Annotating Metadiscourse in Academic Lectures from Different Disciplines. *Ghada Alharbi, Raymond W. M. Ng and Thomas Hain (2015)*, *In the International Workshop on Speech and Language Technology in Education (SLaTE)*, 2015

5. The OpenCourseWare Metadiscourse (OCWMD) Corpus. *Ghada AlHarbi and Thomas Hain*, *In the 10th Edition of the Language Resources and Evaluation Conference (LREC)*, 2016

6. Metadiscourse Tagging with Convolutional Neural Network. *Ghada Alharbi and Thomas Hain (In preparation for the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017

7. Metadiscourse Tagging in Academic Lectures. *Ghada Alharbi and Thomas Hain (In preparation for the IEEE Transactions on Learning Technologies)*, 2017

Chapter 2

Metadiscourse Tagging Approach in Academic Lectures

The focus of this thesis is to define and develop a complete system for metadiscourse tagging of academic lectures, and then validate the automatically detected tags using a thematic discourse segmentation task. This system consists of four-stages; it is crucial to understand the purpose of each stage individually, and then the relationships between them.

The purpose of this chapter is to introduce the metadiscourse tagging approach and some details of the four stages, as well as how some of them contribute to solving the problem of understanding lecture discourse structures. It also reviews some work relating to corpus building and sentence-level modelling. Additionally, it describes the information that underpins the work presented in the following chapters, including the lectures datasets used throughout the work and the application of thematic discourse segmentation, developed to evaluate the adequacy of these tags in recovering the high-level structures of lecture discourse. More technical details for each stage are presented in the following four chapters.

2.1 Introduction

Metadiscourse tagging is often required for discourse-based applications that approach discourse understanding by assigning functions at sentence/utterance-level. This is because metadiscourse can serve as a guide in the communication events which help student to understand the information given. Most studies on metadiscourse tagging have a common strategy that involves two important components: corpus building and model development. The former is needed as metadiscourse tagging tasks target different domains and for different applications and it is rarely find a corpus that serves one interest. For that purpose, a

specific annotation scheme needs to be defined that is suitable for both the target data type and the target application. Interestingly, most of metadiscourse tagging studies have also a common strategy in the annotation methodology, in which sentences are assigned functions based on the semantic meaning of specific phrases (Correia et al., 2014b, Teufel, 1998). These functions are drawn from pre-defined categories in the annotation scheme. Further, the development of the metadiscourse tagging model involves the features set and the classification algorithm. These two components are key in developing a robust tagging model that is capable of dealing with the problem of metadiscourse expressions variants.

From these observations in previous works, it is clear that three main decisions need to be taken when developing metadiscourse tagging for a new domain (*e.g.,* academic lectures). Firstly, one needs to either define or adapt an existing scheme that is suitable for both the data type and the end application. For example, Teufel (1998) defines an annotation scheme that is adequate for scientific articles, while Correia et al. (2014b) adapt an existing one that is designed for academic speech. Secondly, finding feature sets that are more representative for the task. For example, both Teufel (1998) and Correia et al. (2014a) utilise word n-grams features due to their ability to capture the metadiscourse expressions in the sentence. However, word n-grams suffer from a sparsity problem as it needs a huge amount of labelled data to be able to cover all expressions variants of a particular metadiscourse category. For this reason, Correia et al. (2014a) add other features such as Part-of-Speech (POS) tags and sentence positional information in order to improve the model performance. Finally, another decision needs to be made with regards the classification algorithm that is able to utilise the set of features to boost the model performance in identifying and classifying variants of metadiscourse expressions.

In this thesis, an analysis of metadiscourse in academic lectures was carried out by investigating the adequacy of the chosen metadiscourse scheme for academic lectures from two different disciplines, Physics and Economics. This analysis allowed the building of a corpus of metadiscourse in academic lectures. Then, an automatic speech recognition (ASR) was built for lectures, in order to evaluate the performance of metadiscourse tagging models on these automatic transcripts. However, to complete the evaluation process on such transcripts and due to the lack of time-stamps in the reference transcriptions, an alignment process was applied to project the gold-standard metadiscourse tags from the reference transcripts onto the automatic ones. Subsequently, the problem of assigning metadiscourse tags for each sentence in the lecture was defined as a multiclass classification task at sentence level. First, a set of text-based and acoustic-based features was defined that had significant effects on closely related tasks. These features were then fed into an SVM classifier able to deal with both high-dimensional space features (text-based) and low-dimensional space features (acoustic-based), to allow more compact features representations. Finally, an alternative model was defined as well, based on CNNs and continuous representations of features sets, including

word embeddings. The purpose of the latter model was to advance the discovery of the underlying structures of ordered sequences of words, such as metadiscourse expressions, and hence boost the classification performance.

The following sections review recent works in building discourse-annotated corpora, and provide some details about the approaches developed, along with the lectures datasets used for the development of the metadiscourse tagging model, and the applications used for evaluation.

2.2 Related Work

This section provides an overview of existing discourse corpora that address either discourse functions in general, or metadiscourse in particular. It is worth noting that this does not give complete coverage to all research focused on discourse or metadiscourse in English. Instead, the aim is to report resources related to the task of identifying metadiscourse, particularly in spoken language. Important modelling approaches regarding metadiscourse or similar phenomena are also briefly reviewed, since these studies are related to the tagging models described in Chapter 5.

2.2.1 Discourse-annotated corpora

Discourse Treebank (RST-DT)

RST-DT is a discourse-annotated corpus developed by Carlson and Marcu (2001) to be used by the NLP community. It consists of a collection of *Wall Street Journal* articles taken from the Penn Treebank. The corpus is based on a semantics-free theoretical framework of discourse relations proposed by Marcu (2000) for text summarisation, but it is general enough to be used for any application that requires discourse analysis. This framework was based on Rhetorical Structure Theory (RST), as defined by Mann and Thompson (1988). RST is one of the best-known discourse analysis frameworks. In this framework, a discourse tree can be used to represent a coherent text, in which its leaves represent non-overlapping text fragments, which are referred to as elementary discourse units (EDUs). Then, adjacent text nodes may establish relations with other adjacent nodes in the tree structure. Relations can be of an intentional, semantic, or textual nature. More formally, this corpus contains 24 discourse relations, which are further divided into a set of 16 relation classes, with a total of 78 finer-grained rhetorical relations. A simplified version of this structure under RST is presented in Figure 2.1.

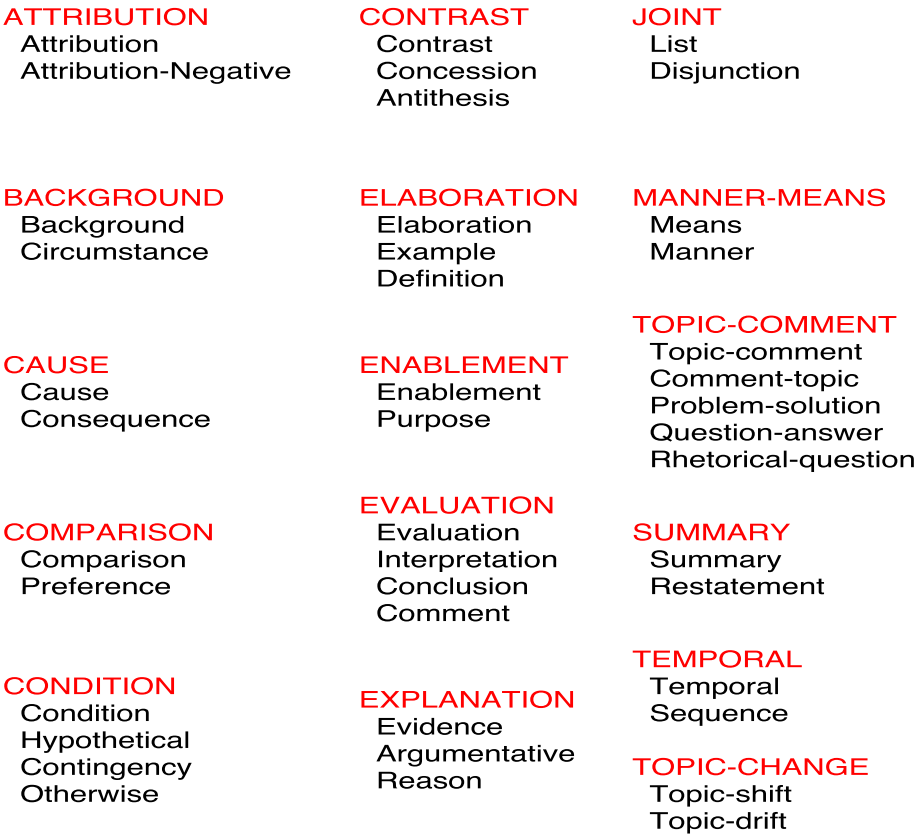


FIGURE 2.1: Simplified Rhetorical Structure Theory categories, adapted from Carlson and Marcu (2001).

Most of the categories in the RST-DT structure are intended to find rhetoric relations between two EDUs. However, some of them intersect with the objectives of a metadiscourse functional approach, in signalling the discourse function. In other words, some of the categorises in the RST-DT match the definition of similar categories in the metadiscourse annotation scheme, which will be further demonstrated in Table 3.1 in Chapter 3. For example, the category Example in the RST-DT matches *Exemplifying* in the metadiscourse annotation scheme. Similarly, the category Restatement matches *Reformulating and Clarifying* in the metadiscourse scheme, as does *Definition* in the RST-DT with *Managing Terminology*.

The RST does not require some expressions in the text unit to highlight a relationship, as is the case with metadiscourse. However, there are some cases where the existence of some cue phrases, such as "but" and "because" indicate a discourse relation. For example, it has been noticed that in the RST-DT corpus only 61 of 238 Contrast relations and 79 out of 307 *Explanation-evidence* relations were marked by a cue phrase (Marcu and Echihabi, 2002).

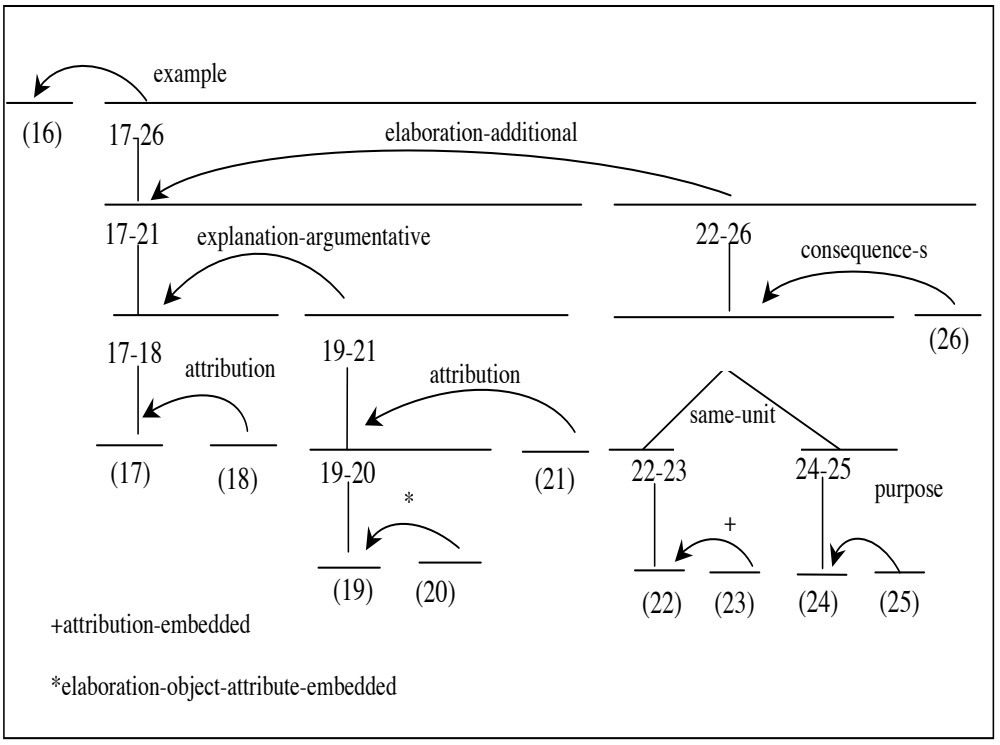


FIGURE 2.2: RST discourse sub-tree for multiple sentences, adapted from Carlson et al. (2003).

Penn Discourse Treebank (PDTB)

PDTB is another annotated discourse relation corpus built upon Penn TreeBank, proposed by Marcus et al. (1993), a corpus well-known in the NLP community for training parsing models. It consists of excerpts from the Wall Street Journal (WSJ). PDTB is considered the largest manually annotated discourse relations corpus to date, and was developed by Prasad et al. (2008). PDTB is not based on the framework of RST, as is the case with RST-DT. Instead it follows the framework presented by Webber (2004) of a predicate-argument framework with a different set of predefined discourse relations.

Unlike RST, PDTB requires the existence of discourse connectives, such as "because" to determine whether there is a relation between two text spans. Discourse connectives can be a word, a phrase, or a pair of phrases whose interpretation conveys a semantic relationship between two abstract objects (Asher, 2012). This type of discourse relation is called an explicit relation, due to the presence of discourse connectives in the clause, and can be organised according to four syntactic categories (Webber et al., 2005):

- **Subordinating conjunctions** – *e.g.* because, although, when, if, as;
- **Coordinating conjunctions** – *e.g.* and, but, so, nor, or;

- **Subordinators** – *e.g.* provided (that), in order (that), except (that);
- **Discourse adverbial** – including adverbs, *e.g.*, instead, therefore, and prepositional phrases, *e.g.* on the other hand, as a result.

However, discourse connectives can have more than one meaning, and determining the correct meaning of connectives is important for several discourse relations tasks. Consequently, Miltsakaki et al. (2008) intended to identify the sense of such discourse connectives by organising them into 4 categories, which can be further divided into 16 types and 23 subtypes (see Figure 2.3). This highlights another important difference between RST and PDTB, which is that PDTB does not have a tree-style structure when coding its relations types over discourse sentences (see Figure 2.2 on how that be done for the RST case). PDTB organises them hierarchically. Despite the fact that the structure of PDTB works on low-level discourse contexts, it does not represent the functional aspects of the discourse sought in building this study's corpus of metadiscourse in lectures.

Mentioned Language (ML)

Some approaches to metadiscourse study the notion of metasemantics, which is defined as the use of language to analyse and describe semantics. Such use of language is often referred to as **use-mention**, and was originally presented by Lyons (1977) to distinguish between the usage of words or phrases in two particular cases:

- **Use** – use of language where words are mapped to concepts from outside the language; *e.g.* I watch **basketball** at weekends.
- **Mention** – use of language in which it is not the concept that a word represents, but the word itself; *e.g.*, The term basketball may denote one of various sports.

The first approach proposed by Wilson (2010) to develop a scheme for mentioned language sought to annotate a corpus consisting of 1339 sentences. The categories of this scheme, along with examples for each of them, are shown in Table 2.1. Note that each category in this scheme is named according to the language it is pursuing, such as translations, phonetics and symbols.

In a related study by Wilson (2012) the previously proposed set of metasemantics categories is refined using the *English Wikipedia*[1] corpus. In particular, the study utilised previous knowledge composed of a set of 23 nouns and verbs that can be used as indicators of mentioned language:

[1] http://en.wikipedia.org/wiki/EnglishWikipedia

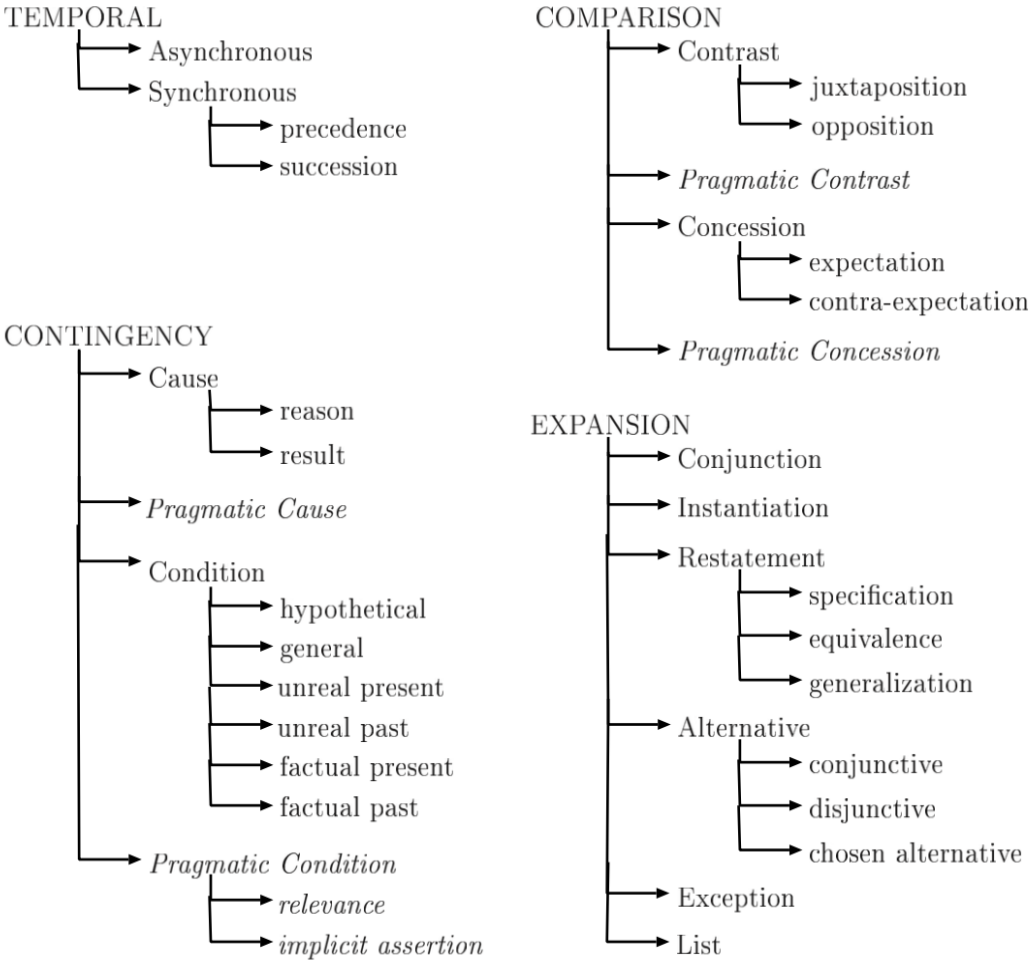


FIGURE 2.3: Hierarchy of Penn Discourse Treebank (PDTB) sense tags. Taken from Miltsakaki et al. (2008),

- **Nouns** – *e.g.*; letter, meaning, name, phrase, pronunciation, sentence, sound, symbol, term, title, word.
- **Verbs** – *e.g.*; ask, call, hear, mean, name, pronounce, refer, say, tell, title, translate, write.

Then, these group of words were used as hooks to retrieve a set of candidate sentences that matched the notion of mentioned language from the Wikipedia corpus. This set of candidate sentences was then assigned one of the following categories:

- **Words as Words (WW)** – the phrase is used to denote the word or phrase itself, this is similar to the category Words as themselves in Table 2.1;
- **Names as Names (NN)** – this captures uses of a phrase as a proper name; again this is similar to the category of Proper name in Table 2.1;

Category	**Example**
Proper Name	A strikingly modern piece called *The Pump Room*
Translation	The Latin title translates as *a method for finding curved lines*
Attributed Language	*I read a chess book of Karpov*, the 21-year-old said
Words as Themselves	*Submerged forest* is a term used to describe the remains of trees"
Symbols	He also introduced the modern notation for the trigonometric functions, the letter *e* for the base of the natural logarithm
Phonetic	The call of this species is a high pitched *ke-ke-ke*
Spelling	*James Breckenridge Speed* (middle name sometimes spelled *Breckinridge*)
Abbreviation	often abbreviated *MIIT* for *Moscow Institute of Transport Engineers*

TABLE 2.1: Wilson's taxonomy of mentioned language, along with some examples of each. *Italics* refer to the mention, and underline text denotes the use of the language.

Category	**Overall Occurrences**	**Sample Occurrences**	κ
WW	438	17	0.38
NN	117	17	0.72
SP	48	16	0.66
OM	26	4	0.09
XX	1764	46	0.74
Total	2393	100	

TABLE 2.2: The annotation results of Wilson (2012). κ refers to the agreement metric used.

- **Spelling or Pronunciation (SP)** – text that describes a spelling or pronunciation; it shares some similarity with the category Spelling in Table 2.1;
- **Other Mention (OM)** – this refers to an instance of mentioned language not fitting the above categories;
- **Not Mention (XX)** – candidate instances but not a representative of *mentioned language*.

After this classification process, the next step was to label the subset of 100 candidate instances by hiring three expert annotators, who were given guidelines that also included the above five categories. Agreement between annotators was measured using *Fleiss' Kappa* coefficient κ. Table 2.2 shows the number of occurrences in the retrieved set, per category annotated, by the first author of the study, alongside the correspondence frequency in the set of 100 candidate instances, and with the κ coefficient.

The analysis of both retrieved candidate sets revealed that only 26% were annotated by the first author of the study as *mentioned language*, and not *mentioned language* comprised about 1,764 out of 2,393 total instances. As regards the 100 sample, it showed that the expert annotators had no problem in classifying an instance as mentioned language or not, as indicated by the reported κ of 0.74. However, annotators faced some difficulties in classifying metalanguage according to the pre-defined categories, with κ between 0.09 and 0.72. This

ADD	include both Adding to Topic and Marking Asides
ANT	Anticipating Response
ARG	Arguing
CLAR	Clarifying
COM	Commenting on Linguistic Form/Meaning
CONC	Concluding
DEF	Definitions (originally, Manage Terminology)
DELIM	Delimiting Topic
EMPH	Emphasizing (originally Managing Message)
ENUM	Enumerating
EXPL	include Exemplifying and Imagining Scenarios
INTRO	Introducing Topic
POST	Postponing Topic (originally, Previewing)
RCAP	Recapitulating (subdivision of Reviewing)
REF	Refer to Previous Idea (subdivision of Reviewing)
R& R	collapsed from Repairing and Reformulating

TABLE 2.3: The annotation scheme used by Correia et al. (2016), which has been adapted from Ädel (2010).

indicates that annotators tend to agree about whether the given candidate is *mentioned language* or not, but not about classifying them according to their function.

Metadiscourse in TED Talks (metaTED)

The approach of Correia et al. (2016) to studying metadiscourse in spoken language is the most closely related to the current study, because of the scheme used to signal the discourse functions with a wide range of functional categories. Although Correia et al. (2016) claims that the results of his study are beneficial for a major goal, the task of making a presentation skills instruction tool, the scheme used is general enough to apply to any natural text, and concise enough to offer an algorithmic approach to discourse analysis. This scheme was originally proposed by Ädel (2010), with 23 specific categories and 4 generic ones. It is the same scheme that was followed to build the metadiscourse in academic lectures corpus further described in Chapter 3. Unlike this study's approach, however, the study recruited a crowdsourcing service and applied a quality assurance mechanism to build this corpus using TED talks.

As a first approach, Correia et al. (2014b) annotated 180 TED Talks with 4 metadiscourse categorises from Ädel's scheme. These categories are: *Introducing Topic*, *Concluding Topic*, *Marking Asides*, *Exemplifying*, and *Emphasising*. The results of this first study indicated that the crowd were able to annotate the functions of metadiscourse on presentation-style talks such as TED Talks, and at the same time provide detailed analysis of the level of

complexity, with different categories in the scheme. The initial study thereby attempted to assess the understanding of the crowd through these categories.

In follow-up studies (Correia et al., 2015, 2016) the authors attempted to expand the categories of the Ädel (2010) scheme by having 16 tags in total, as shown in Table 2.3. As a result, a new annotated corpus was developed called *metaTED*, using the same collection of TED Talks as previously, covering a wide range of general topics. As in the previous study, the annotations were done using crowdsourcing, but a small set thereof was validated by the experts. Agreement of results indicated different levels of understanding among the crowd with regards to metadiscourse categories, in the range of [0.15-0.49]. Similar behaviour was also noticed with the agreement obtained from the experts [0.18-0.72]. Based on these low agreement results, the study concluded that only 10 categories out of 16 could be used in further NLP tasks.

Although the work in this thesis relies on the same metadiscourse scheme used by Correia et al. (2016), the target dataset is different, as this study examines the phenomena in terms of university lecture courses that contain a set of related lectures and topics from the OCW platforms. This type of data fits better with the scheme proposed by Ädel (2010), as it was also the target dataset in the small analytical experiment in that study, as will be further illustrated in Section 3.2, Chapter 3. For instance, the categories Previewing and Reviewing are more appropriate to be used in analysis of lecture courses than in random talks presented via the TED platform. Besides, this study applies some mechanisms that facilitate the annotation jobs, in turn increasing annotators' level of understanding of the task. Moreover, the use of expert annotators with knowledge of both the subject matter and the metadiscourse increased their level of agreement and consequently the frequency of these categories in the gold dataset. All of these factors suggest that metadiscourse in academic lectures is naturally frequently occurring, and that building a classifier to automate this process can be beneficial for a number of NLP-tasks, such as thematic discourse segmentation, which will be studied in the application chapter (Chapter 7).

2.2.2 Modelling Approaches

Manually annotating metadiscourse within any new corpus can be very costly. For that reason, efforts have been made to develop automatic methods for metadiscourse tagging and discovery. Most of the proposed works in the literature develop a predictive paradigm, where metadiscourse models are first trained on a small corpus and afterwards used to predict unseen sentences. However, there are other studies that approach the task using an unsupervised paradigm. The modelling approaches to metadiscourse can thus be supervised or unsupervised, but the main focus here will be on supervised approaches.

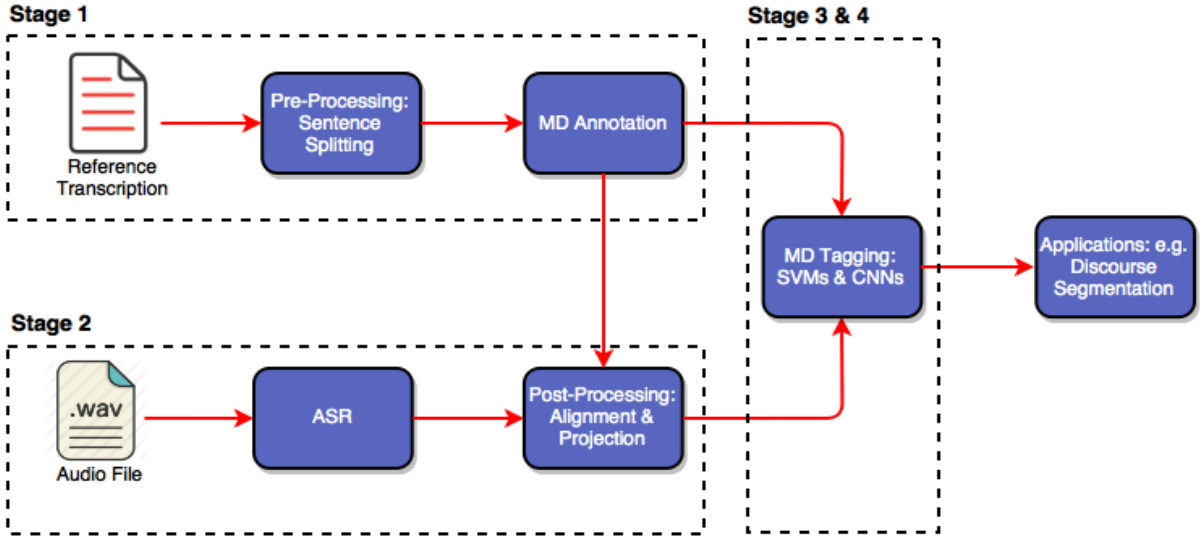


FIGURE 2.4: The approach for metadiscourse tagging. There are four stages: metadiscourse annotation using reference transcriptions as input; generating automatic transcriptions with audio file as input; metadiscourse tagging model with SVMs; and metadiscourse tagging model with CNNs. Note that stage 1 is the first stage, as the output of this is used to produce the output of stage 2 (*i.e.,* ASR outputs with corresponding gold standard metadiscourse tags). Stages 3 and 4 are implementing the same task but with different features and classifiers for a purpose of comparison and improvement.

The metadiscourse modelling schemes that are commonly used for metadiscourse tagging are traditionally chosen from the same set of general machine learning methods used in most natural language processing tasks, such as Decision Trees (Correia, 2013, Teufel, 1998, Wilson, 2013), Naive Bayes (Wilson, 2013), Support Vector Machines (Wilson, 2013), and Conditional Random Fields (Madnani et al., 2012). Regarding features, most metadiscourse tagging models rely mainly on lexical features; these include word n-grams and POS tags. Metadiscourse history and positional information are also often used as relevant information. However, previous studies of metadiscourse tagging have not explored the effects of prosodic cues on the task. This issue and more technical information about these approaches are discussed in further detail in Section 5.2.

2.3 Approach Description

Understanding and analysing academic lectures discourse is at the heart of most education-based applications. Assigning a discourse function to a particular word based on the context, such as discourse markers (*e.g.* 'however'), is one method working towards discourse understanding. Another method is metadiscourse tagging which assigns a discourse function to a sentence (or utterance) based on the presence of certain expressions. The annotations scheme that defines the set of discourse functions for the sentences is often domain-specific in design.

For example, the set of metadiscourse tags that can be applied to scientific articles is very different from those applied to academic lectures.

Previous work on closely related definitions of metadiscourse for spoken discourse has focused on applying the annotations scheme to a set of TED Talks. However, this method did not investigate whether disciplinary knowledge has any effect on the annotation study. It also relies only on a set of text-based features in the modelling process, such as n-grams features. Therefore, there is a need to first explore the effects of discipline knowledge on the metadiscourse, by applying it to a more challenging dataset – in this case university lecture courses from two different disciplines: Physics and Economics. In addition, the method does not exploit the usefulness of spoken language artefacts such as prosodic features for the modelling approach. In this thesis, a four-stage approach is proposed to provide a robust model for metadiscourse tagging in academic lectures, using feature combinations from two modalities, text-based and acoustic-based, and two different classifiers for the purpose of comparison. The architecture of this approach is shown in Figure 2.4, and the general ideas underpinning each stage are given below.

Stage 1: A Corpus of Metadiscourse in Academic Lectures

In order to understand and analyse lecture content at sentence-level, the metadiscourse tag assigned to a particular sentence needs to signal its discourse function. Therefore, the corpus used needs to satisfy two conditions: being designed for the lectures domain, and having categories that describe the purpose of the lecture sentences. This is important because these functions can later help to interpret the high-level structure of the lecture content, or to serve as features for subsequent applications, such as summarisation and tables of content for web browsing. However, such conditions are hard to find in the existing discourse-annotated corpora (as discussed in Section 2.2.1) as most of them were built specifically for written discourse and do not reflect the discourse function at sentence-level; those few that are designed for spoken discourse and present the function of the discourse are not designed specifically for academic lectures.

To solve this problem, the first step in implementing this metadiscourse tagging approach was to build a corpus of metadiscourse in academic lectures. This was achieved by using an annotation scheme designed to express the discourse functions at sentence-level, such as the one proposed by Ädel (2010). This study has proved to be an effective scheme for that purpose, in annotating metadiscourse in spoken language using presentation-style datasets such as TED Talks as reported by Correia et al. (2014b). Extending this scheme to academic lectures is beneficial for various education-based applications. Moreover, the scheme allows grouping of metadiscourse categories at two levels of granularity: generic (contains

4 categories) and specific (contains 19 categories). Each one of the 4 generic sub-groups can serve a general discourse function: *Metalinguistics Comments*, *Discourse Organisation*, *Speech Acts*, and *Interaction with Audience*. Providing both levels of metadiscourse allows one to report and compare the performance of the tagging model using both generic and specific metadiscourse categories. After determining the appropriate scheme, the next step is to choose the set of lecture courses from the OCW platforms, which are freely available online, and will be described in further detail in Section 2.4. However, such datasets required some preprocessing, for example by splitting manual transcripts into sentences using a sentence splitting tool, as shown in the upper panel of Figure 2.4. Due to the complexity of the lecture content, subject expert annotators were hired to conduct the annotation experiments, rather than relying on a crowdsourcing service as demonstrated in Chapter 3.

Stage 2: Automatic Transcriptions of Academic Lectures

The annotation task uses reference transcription to generate metadiscourse tags for each sentence. Producing the reference (or manual) transcription is a time-consuming task, and is sometimes prone to errors, such as when the transcribers are not aware of the technical words used and record other terminology instead (Hazen, 2006). For this purpose, previous works have usually used an ASR system for lectures, such as that presented by (Glass et al., 2007), which provided automatic transcriptions. The two main components in any ASR system are the acoustic model (AM) and language model (LM). The performance of such models suffers from mismatch problems, either in the AM or LM. This usually arises when the test set is different from the training dataset. Various adaptation techniques for both models have been proposed in the literature to reduce the mismatch problem. Deciding which one to follow depends on the task and the speech data type.

For these reasons the second stage develops an ASR system for OCW academic lectures, to produce high quality automatic transcriptions. The system focuses on LM adaptation using a linear interpolation technique, by combining both in-domain and out-of-domain materials. The in-domain resources are derived from a set of academic lectures from a wide-range of disciplines, selected specifically to be similar to the target lectures, in order to reduce the effects of the mismatch problem. The out-of-domain are a large collection of written resources extracted from the web. This approach performed remarkably better than other adaptation techniques. It is also considered simple, fast, scalable and competitive. To improve the results further, the AM was trained using in-domain datasets and the model is based on deep neural network (DNN) in direct combination with a Hidden Markov Model. Such models usually outperform the HMM-GMM systems for the task of large vocabulary speech recognition (LVSR) (Dahl et al., 2012, Hinton et al., 2012, Seide et al., 2011). The ASR system of this stage is evaluated using the standard word error rate (WER). However, the

reference transcriptions lack time-stamp information, which is important for completing the evaluation process. For this reason, a lightly supervised alignment model is applied, which has the benefits of both correcting some errors in the reference transcriptions, and providing time-information for scoring the ASR output. This stage is outlined in the bottom panel of Figure 2.4.

Stage 3: Exploring Features for Metadiscourse Tagging with SVMs

Sentence modelling has been introduced to solve many problems in both NLP and spoken language understanding tasks, including metadiscourse tagging. The aim of a sentence model is to analyse and represent the meaning of the sentence for either classification or generation. To achieve that, one needs to represent sentences in terms of features capable of capturing such meaning. That is, the fundamental step in a sentence model is to find a feature set that is representative for the task, then feed these features to the selected classifier. Various feature types have been proposed for this purpose. Previous work on metadiscourse tagging has proved the effectiveness of using n-grams features for the task (Correia et al., 2014a). There are other feature types, such as prosodic cues, that have not been used before for metadiscourse tagging but usually improve the classification performance in a number of sentence-level classification tasks, such as dialogue acts tagging (Shriberg et al., 1998, Stolcke et al., 2000). However, for the task of metadiscourse tagging it is not clear whether the inclusion of prosodic features would be complementary to textual features, whether it would improve the model performance, or have no effect at all.

Therefore, the third stage of the metadiscourse tagging approach addresses the baseline tagging model, by exploring two kinds of feature sets: text-based and acoustic-based. Finding the best combination of feature sets for the task is a primary focus in this thesis, due to the significant impact of the choice on the model performance. In order to combine high-dimensional (*e.g.,* word n-grams) with low-dimensional features (*e.g.,* prosodic cues), a support vector machine (SVM) was used, which allows easy integration of both modalities (Joachims, 1998). This is because SVMs can learn independently of the dimensionality of the feature space, by measuring the complexity of hypotheses based on the margin with which they separate the data points, and do not depend on the number of features. This means that one can generalise even with very sparse high-dimensional features, such as textual features. To check the robustness of the developed model, a number of test cases were set out for that purpose, including generic and specific metadiscourse tags, and ASR output. This model was evaluated using commonly known metrics – precision, recall and F-1. This stage is placed in the middle panel of Figure 2.4.

Stage 4: Improving Metadiscourse Tagging with CNNs

The traditional approach of combining several hand-engineered features with SVMs has been introduced in the previous stage, to develop a model for metadiscourse tagging. The model showed the ability to classify metadiscourse tags in sentences, and the most effective features for the task were n-grams features, especially word n-grams. The main drawback of using a model with such features is the sparsity problem, as the task of metadiscourse tagging relies on the existence of certain expressions in the sentence. They require a large labelled dataset to cover all the different variants of metadiscourse expressions. In other words, the model will not be able to generalise well for unseen n-grams. This can be solved by using Continuous Bag-of-Words (CBOW; Mikolov et al. (2013)), which maps words to high-dimensional vector representations able to capture both the syntactic and semantic similarities between words, and thus solve the generalisation issue. A downside of CBOW is that it ignores the word order completely, something which is very important to maintain when classifying metadiscourse tags.

To alleviate this shortcoming, the final stage uses a Convolutional Neural Network (CNN) for the metadiscourse tagging task. CNN with a single hidden layer has proved to be an effective model for classifying of sentence-level in a number of NLP tasks (Kim, 2014). CNNs are designed to identify the most inductive features of an ordered set of items locally, regardless of their position in the sentence, for the classification task at hand. A key aspect of using a CNN model is to represent the features using dense, low-dimensional vectors, instead of sparse, high-dimensional vectors. Another useful characteristics of the continuous vectors used is their generalisation power, as similar words have similar vectors. However, it does require large amounts of labelled data to train the network to produce such continuous vectors, and the metadiscourse tagging task has limited amounts of labelled training data available. Thus, these continuous features were obtained from pre-trained word embedding vectors (*e.g.,* word2vec) which can then be tuned for the task. This model is evaluated using the same metric and test setup as in the previous model. This part of the metadiscourse tagging approach is placed in the right panel of Figure 2.4.

2.4 Source of Lectures Data

The first step in developing and evaluating the metadiscourse tagging approach for academic lectures is to select a source of data appropriate for lecture discourse analysis. One of the goals of this thesis is facilitating access to the OCW lecture platforms that are freely available online, permitting use of high quality educational materials organised as courses

under creative commons licences[2]. This is because the creating and preparing of these OCW online courses requires substantial initial and ongoing investments of human labour. Unlike other online courses, such as massive open online courses (MOOCs), the OCW content is less structured, which clearly means there would be a benefit to an automatic process for organising materials, to aid the learning process. For these reason, we restrict the analysis to OCW sources that meet the following criteria:

1. could be found on a wide range of topics of related lectures for two different disciplines, *i.e.* lectures courses, and provided by different speakers within each discipline, in order to have a representative set;

2. provided audio material, which will be used in the development of the ASR system, to generate automatic transcriptions;

3. provided reference transcriptions that are useful for the annotation task;

4. provided segments boundaries that represent the discourse structure of the lecture, which is useful for the application task, to validate the proposed metadiscourse tagging approach.

There are many OCW sources of spoken discourse from different universities, such as MIT OCW[3] at the Massachusetts Institute of Technology, Open YALE Courses[4] at the University of YALE, UCI OCW[5] at the University of California, Irvine, and Stanford OCW[6]. However, not all of them fulfilled the aforementioned criteria. The comparisons between these resources soon led us to choose both MIT OCW and Yale OCW over the other universities' platform. Firstly, MIT and Yale OCW are the only platforms providing the gold standard of discourse segment boundaries, which will be used to train and test the application task in this thesis. Secondly, MIT and Yale are known to provide high-quality recordings and use the same settings across all lectures, which is beneficial in developing the ASR system. This contrasts with other OCW platforms, which used different recording conditions, making them less easy to process automatically.

Another decision was necessary regarding the variety of disciplines to choose from in formulating the final dataset. Lecture courses from two different disciplines – Physics and Economics – were chosen. This decision was mainly based on the availability of lecture resources of similar introductory courses taught across the two different platforms, MIT OCW

[2] https://creativecommons.org/licenses/by-nc-sa/3.0/us/
[3] http://ocw.mit.edu/index.htm
[4] http://oyc.yale.edu
[5] http://ocw.uci.edu
[6] http://online.stanford.edu/courses

	Physics	**Economics**	**Overall**
# Lecture	57	49	106
Average # of segments per lecture	6	7	6.5
# Segment	395	354	749
# Token	4004990	3894639	7899629
# Words	11309	14280	25589
# Utterance	32903	30756	63659

TABLE 2.4: Lecture Corpus Statistics. The first column shows the statistics for the collection of Physics lectures, in terms of average number of thematic segments per lecture, number of thematic segments, and numbers of tokens, words and utterances, respectively. The second column presents similar statistics for the set of Economics lectures. The last column presents the overall statistics across both disciplines.

and Open YALE Courses. For example, the Physics course from MIT OCW is called "Classical Mechanics" and the one from Open YALE Courses is called "Fundamentals of Physics". These courses cover approximately the same scientific material but they are taught by different lecturers from the different institutions. Another reason for choosing these disciplines is to enable an investigation of whether there is a difference in detecting metadiscourse acts between Natural Science and Social Science lectures. In total, 106 OCW lectures were collected from the two disciplines (2 courses for each); the following section provides more detail about the chosen courses.

Physics Lecture Dataset

The Physics dataset consists of spoken lecture transcripts taken from an undergraduate introductory Physics class. Two Physics courses have been included, one from MIT OCW, and the other from YALE Open Course. In contrast to the Economics lectures datasets, this corpus contains much longer texts and consists of 57 lectures. A typical lecture of 75 minutes has 500–600 sentences, with up to 8500 words in each, which corresponds to about 15 pages of raw text. Table 2.4 shows further statistics for the Physics corpus.

As stated above, this corpus also contains annotations for thematic segment boundaries, with segments labels. The thematic segments herein are, in fact, a multi-dimensional, heterogeneous collection of pragmatic and semantic-oriented text units. These thematic segmentations were produced by the teaching staff of the Physics course at MIT and Yale. As stated earlier, the objective of these materials was to facilitate access to lecture recordings available on the class website under the OCW initiative. On average, a lecture was annotated with six segments, with a typical segment corresponding to two pages of a transcript. These segmentation boundaries are required later, to investigate whether metadiscourse tags are indicative of high-level discourse structures (thematic boundaries), as will be demonstrated in Chapter 7.

Economics Lecture Dataset

The second lecture dataset differs in both subject matter and lecturing style. This dataset comprises two economics courses taken from MIT OCW and YALE Open Course. The undergraduate introductory economics corpus has, in total, 49 lectures of 75 minutes and, on average, 650–800 sentences per lecture, which corresponds to roughly 8500 words. Further statistics about the Economics lectures are also presented in Table 2.4.

As with the Physics lectures, the thematic segmentation boundaries were obtained from the course website, and again, objective of these lectures is to facilitate access to OCW resources. On average, an Economics lecture was annotated with seven segments, with a typical segment corresponding to two pages of transcript. As was noted with the Physics lectures, the thematic segmentations were a heterogeneous collection of pragmatic and semantic oriented units. The thematic boundary annotations are used as the gold standard for the application task (thematic segmentation) and are presented in this thesis in Chapter 7.

2.5 Application

Providing a discourse analysis tool for enriching the lecture discourse with functional tags such as metadiscourse can be a basic step in several tasks, such as summarisation, search and retrieval, indexing, and browsing. Successful applications in the development stage that succeed with a challenging dataset such as academic lectures give confidence that the presented approach can be practical for various education-based applications. Metadiscourse has proven to be effective in various applications, for example: summarising a meeting according to its activities Niekrasz (2012), modelling argumentative zoning for scientific research articles (Teufel and Moens, 2002), and most recently, designed to be used in building presentation skills tools, using TED Talks (Correia et al., 2014b).

The motivation: The usefulness of the metadiscourse tagging approach in this thesis is verified through its application in finding lecture discourse structures. This task is referred to as the thematic segmentation of academic lectures, and aims to segment lecture into either topical or functional themes. The main intuition behind this is driven by analysing the manual annotation of the thematic segmentation, which reveals that a mixed approach of semantic and pragmatic elements is needed, as demonstrated in Section 1.1. The same observation of an approach combining semantic and pragmatic segment boundaries for meetings discourse was made by Niekrasz (2012).

The task: The thematic segmentation of academic lectures is defined as a binary classification task. Then, the annotated metadiscourse tags that were detected either manually or

automatically using the previous models are combined with lexical cohesion base features, for the task of the thematic segmentation of the lectures. Lexical cohesion is defined by Morris and Hirst (1991) as:

the result of chains of related words that contribute to the continuity of lexical meaning. These lexical chains are a direct result of units of text being about the same thing, and finding text structure involves finding units of text that are about the same thing.

The lexical cohesion score was extracted using the LCSeg algorithm, which computes a lexical cohesion score by combining information from all term repetitions at each the end of each utterance (Galley et al., 2003).

2.6 Summary

This chapter has presented a general overview of the metadiscourse tagging approach being developed, the proposed model components, the annotated corpus behind this approach and other experiments defined throughout the thesis. A more in-depth description of each stage is provided in the following chapters. Chapter 3 covers the first stage, related to metadiscourse corpus building in academic lectures from two different disciplines, Physics and Economics. Chapter 4 presents the system of automatic transcription of academic lectures as a second stage. Chapter 5 is related to features extraction and the baseline model of the metadiscourse tagging stage, while Chapter 6 is related to the neural network based multi-class classification model stage. An application related to the thematic segmentation of academic lectures task and used to evaluate the approach is also introduced in this chapter, and further details of the evaluation results form the content of Chapter 7.

Chapter 3

Annotating Metadiscourse in Academic Lectures

Chapter 2 presented an overview of the metadiscourse tagging approach consisting of four stages, mostly targeting different tasks. This chapter presents the first of these stages, which involves building a corpus of metadiscourse within academic lectures. Recently, *metaTED* corpus was built to study metadiscourse in presentation-style lectures such as TED Talks. The scheme in that study was designed to provide a function-oriented taxonomy of metadiscourse. This existing scheme merged previous approaches to metadiscourse taxonomies, to be applicable to both spoken and written discourse. In this work, the application of this scheme is extended to academic lectures from two different disciplines, Physics and Economics, with the aim of studying the effects of discipline-specific knowledge on the annotation task. This scheme is then further adapted for lectures via a trial study. This adaptation takes into account both the material to annotate and the setting in which the annotation task is performed. Experiments with the selected OpenCourseWare (OCW) lecture datasets described in the previous chapter show that expert annotators are able to identify occurrences of multiple categories of metadiscourse, hence confirming a reliable coding of metadiscourse in academic lectures using the adapted annotation scheme.

This chapter is structured as follows: Section 3.1 introduces the annotation experiments, along with the motivation for this task and its contributions. Section 3.2 reviews previous work related to the metadiscourse schemes, and definitions for its categories from an English language perspective. The implementation methodology of the proposed annotation study is presented in Section 3.3, along with the adapted metadiscourse scheme. The inter-annotator agreement is measured using a commonly known metric – Fleiss's Kappa coefficient – and a summary of the results obtained is given in Section 3.4. Section 3.5 provides a concluding discussion.

3.1 Introduction

Academic lectures, including the OCW online courses, offer rich opportunities for studying a variety of complex discourse phenomena that help in understanding the lecture discourse. For instance, lectures contain regions where lecturers introduce some concepts, emphasise others, interact with students, and engage in other interesting ideas. It is crucial to locate these regions in order for a system to interpret lecture discourse. In other words, the general aim of a system trying to infer a lecture discourse or understand it will be tied to these regions. For example, if the purpose of an utterance is to emphasise a particular concept, the system will thereby be able to determine what information is important to highlight for students, by fetching those utterances labelled as important. In another example, a system may filter utterances based on their functions labels, to recover higher-level forms that might serve as a table of content.

Lectures often contain strategies used to indicate these regions that reveal both topical and functional structure, as well as other high-level discourse dynamics. These strategies are known as metadiscourse and it is used to comment on the language and does not contribute to the content of the lectures (Crismore, 1989). Its play a key role in directing the audience during the lecture, and also in discourse analysis research. Further, metadiscourse can occur in any type of communication, including both speech and writing. For example, Teufel (1998) studies metadiscourse to define structure for scientific articles from two disciplines: linguists and medicine. In speech, Niekrasz (2012) presents a study of metadiscourse to segment meeting conversations based on its activities. More recently, in speech as well, Correia et al. (2016) present a corpus of metadiscourse in TED Talks called *meta*TED which is based on several previous analytical studies discussed in Correia et al. (2015, 2014b) and Correia (2013).

Most of these corpora were built for specific domains and to serve a specific purpose. Finding a metadiscourse corpus for academic lectures that signal the rhetorical functions of the discourse has not yet established. To build such a corpus, several trials of annotation experiments needs to be run, based on a particular scheme. Metadiscourse annotation is defined, therefore, as the activity of labelling the stretches of discourse (*i.e.* expressions) with pre-defined categories, according to a specified scheme that satisfies two conditions: 1. it represents their communicative functions; 2. is suitable for academic lectures. Clearly, finding a proper scheme that meets these criteria is an important step towards building a corpus of metadiscourse.

3.1.1 Motivations

Over the years, a number of metadiscourse annotation schemes have been developed, such as Luukka (1992), Mauranen (2001), Thompson (2003), and Auria (2006). These schemes were all designed for a specific purpose: to serve specific domain or discourse types, such as spoken or written communications. Additionally, most of these metadiscourse schemas focus solely on form rather than function in defining their categories. That said, Ädel (2010) proposed a scheme that defines metadiscourse categories according to discourse functions, and is suitable for both written and spoken discourse. However, the intuition behind this analytical study does not contribute to the goal of corpora building, as it only provides limited examples to support the category organisation decisions.

Although the design of Ädel's scheme was based on an analysis of a collection of academic lecture courses, Correia et al. (2014b) shows the reliability of this model for annotating presentation-style discourse, *i.e.* TED Talks, with non-expert annotators. Naturally, annotation schemes can have the capacity to analyse different disciplines to the one they were designed for, though as Hyland (1998) has argued, metadiscourse varies significantly between research communities, which can be attributed to the fact that metadiscourse has to follow the common standard and expectations of particular cultural and professional communities. Hyland's study shows significant differences in metadiscourse usage across the disciplines of Microbiology, Marketing, Astrophysics and Applied Linguistics. These observations were highlighted for written discourse, in particular scientific articles. However, studying the effect of disciplinary knowledge on academic lectures has not been widely studied.

Building a corpus of metadiscourse for academic lectures from two different disciplines is not just helpful to downstream applications such as improving the performance of thematic discourse segmentation. It also can serve as an opportunity for the related research communities such as the field of English language learning to study the effect of domain knowledge on the phenomenon on a large-scale representative sample built from real academic lectures.

3.1.2 OCWMD-Corpus: Overview

This chapter presents a corpus of metadiscourse for academic lectures from two different disciplines: Physics and Economics. This is achieved by applying a scheme (Ädel, 2010) designed to provide a function-oriented taxonomy of metadiscourse. This existing scheme merged previous approaches to metadiscourse taxonomies to be applicable to both spoken and written discourse. For this study the scheme has been further adapted for lecture courses, based on a trial study. Moreover, a tool was built specifically to conduct the annotation experiments and to ease the task for the annotators. Expert annotators were hired to run

the experiments, due to the complexity of the lectures' content, compared to TED Talks. Inter-annotator agreement was evaluated using a commonly known metric – Fleiss's Kappa coefficient. Experiments with the selected OCW lecture datasets described in the previous chapter have shown that expert annotators are able to identify occurrences of multiple categories of metadiscourse, and hence confirm a reliable coding of metadiscourse in academic lectures using the adapted annotation scheme.

3.1.3 OCWMD-Corpus: Contributions

The main contributions of the proposed work fall into the following categories:

- Application and adaptation of Ädel's scheme to annotate OCW academic lecture courses.
- Construction of metadiscourse corpus in academic lectures, to contribute to natural language understanding, and also useful for improving tasks such as thematic discourse segmentation.
- Investigation of the effects of discipline-specific knowledge on the resulting corpus.

3.2 Related Work

3.2.1 The Theory of Metadiscourse

Metadiscourse is a relatively new and interesting concept, believed to play a key role in language organisation and production. Hyland (2005) defined it thus: "metadiscourse embodies the idea that communication is more than just the exchange of information, goods or services, but also involves the personalities, attitudes and assumptions of those who are communicating". Kopple (1985), meanwhile, defines it as discourse about discourse, and as referring to the speaker/writer's linguistic manifestation to engage with his/her audience. In fact, metadiscourse is closely related to other phenomena such as metatalk, metalanguage, or metacommunication, which are often used to name the language people employ when talking about language. In conclusion, metadiscourse is considered a key way of facilitating communication, supporting the position of the sender (writer or speaker) and establishing a relationship with his/her audience.

However, the research area is not unified with regards to metadiscourse. Instead, there are two relatively different views of metadiscourse, according to Mauranen (1993) and Ädel (2006). The former takes a narrow definition referred to as the "reflexive model", while the

other take a broad definition referred to as the "interactive model". In the reflexive model of metadiscourse, reflexivity in language is stressed and is taken to be the starting point for the category. The idea of reflexivity in language refer to the capacity of natural language to refer to itself (Hockett, 1963, Lucy, 1993, Lyons, 1977). In the other model interaction is the key concept, for example between the speaker and audience.

These two models of metadiscourse attracted a lot of attention in the research community in the mid-80s and form the basis of building metadiscourse schemes, primarily focusing on academic written discourse (Crismore, 1989, Kopple, 1985). Later, in the 90s, some studies addressed these models of metadiscourse in spoken communications, as well. The distinction between spoken and written discourse gives rise to fundamental differences in style and expression between these two types of discourse modalities. According to Biber (1986), writing is more contextualised, elaborated and uses a far more explicit level of expression. Speech, meanwhile, is typically more informal, having more contractions and deletions of relative pronouns, is more interactive and involved, having more occurrences of first and second pronouns; and is also more connected to its physical/temporal context. These differences affect the usage of metadiscourse within spoken language.

The following section therefore looks at metadiscourse as it is used in spoken language. More specifically, it will discuss five existing metadiscourse schemes designed for spoken discourse only, or for both spoken and written discourse, describing and comparing the relevance of the different schemes. Despite the depth and detail of these schemes, it is nevertheless useful to categorise the work according to the type of discourse the annotation schemes are designed to support.

3.2.2 Metadiscourse Annotation Schemes

Luukka (1992)

Luukka proposed a metadiscourse scheme focused on academic discourse that was able to handle both written and spoken discourse (Luukka, 1992). In this study, a small corpus was used, consisting of two versions of five conference papers: the written version of the proceedings, and the transcript of the oral presentation. By analysing the content of this corpus, a taxonomy of metadiscourse was proposed for both speech and written communications. The taxonomy developed differentiated between metadiscourse strategies used for discourse organisation and those used for interaction with the audience; this consisted of three general categories:

- Textual – strategies related to the structuring of discourse.

- Interpersonal – related to the interaction with the different stakeholders (participants) involved in the communication.
- Contextual – covering references to audiovisual materials.

Luukka introduced further sub-functions as part of the proposed metadiscourse taxonomy. The taxonomy is further split on several levels, where the signals of interaction category includes the subfuction 'interpersonal', which may be further refined as i) presence of author (I), and ii) presence of audience (you), and iii) presence of author and audience (we).

Mauranen (2001)

In contrast to Luukka's scheme, Mauranen (2001) focused on spoken materials in developing the proposed metadiscourse scheme. However, Mauranen makes explicit the differences in approaching the metadiscourse phenomena by proposing different taxonomies for each discourse type. In developing this scheme, the Michigan Corpus of Academic Spoken English (MICASE), developed at the University of Michigan's English Language Institute (Simpson and Swales, 2002), was used. This corpus consists of 200 hours of a collection of lectures, seminars and presentations. In contrast with Luukka, who only used monologue discourse, the corpus used by Mauranen contains both monologue and dialogue types of spoken discourse. The scheme devised by Mauranen is composed of three general functions:

- Monologic: structuring of the speaker's own discourse (similar to 'textual' in Luukka's scheme).
- Dialogic – referring to audience interventions or answering questions (similar to 'interpersonal' in Luukka's scheme).
- Interactive – eliciting participation from the audience and manipulating the roles of the stakeholders (also related to interpersonal in Luukka's scheme).

The main observations of these functions indicates that the guiding principle in developing this scheme was based on the speakers involved in the discourse. Some similarities with Luukka scheme can be noted, since both attempt to distinguish between spoken and written discourse in developing the taxonomy. This reflects how students can participate in the spoken discourse in real-time, in contrast to written discourse.

Thompson (2003)

A comparative study of discourse organisation between lectures in academic disciplines and English for Academic Purposes (EAP) classes has been presented by Thompson (2003). The study focused on spoken discourse and aimed to show the difference between EAP courses and real lectures with regards to discourse organisation. For instance, when listening to a lengthy and complex monologue such as an academic lecture, the students formed a mind map of the concepts introduced in the lecture to help them understand the given information. For this reason, the corpora used in developing the scheme are a combination of academic lectures and EAP materials, in order to highlight how frequently real lectures use discourse organisation artefacts, compared to EAP courses. The sample consisted of six undergraduate university lectures and five EAP published listening skills classes. Based on the comparison of these two kinds of materials, a taxonomy of metadiscourse was derived, with three main groups:

- Content Markers: used to give information about the lecture to come.
- Structuring markers: used to outline the structure and sequence of the lecture;
- Metastatements – used to organise the communication event itself (not its content).

Moreover, Thompson attempted to further organise these functions on three levels: global, topical, and sub-topical. That is, labelling each metadiscourse expression based on the level of granularity is a natural property of the communication event as its shows different levels of topics and how the interaction between them is linked in the given lecture.

Auria (2006)

Auria (2006) proposed a scheme that focuses on spoken metadiscourse in academic settings, and contrasts it with both conversational language and written discourse. The study shows that there is an increase in the use of metadiscourse, particularly when the lecturer is attempting to increase comprehension. The author concluded that metadiscourse is an important linguistic resource for analysing academic lectures. Moreover, an increase in metadiscourse occurrences was found in longer academic lectures. This can be attributed to the fact that the demands on a student's attention are greater for a longer lecture, having to remember something said an hour ago instead of twenty minutes ago. For this reason, lecturers try to maintain students' attention by using discourse organisation strategies such as metadiscourse, to increase the level of comprehension.

The main intuition in developing Auria's scheme is based on the intention of the lecturers. The analysis was done using the MICASE corpus to derive a scheme consisting of three categories of metadiscourse:

- I-pattern – expressions that use the first person singular nominative pronoun, such as 'I'm gonna' or 'I wanna'.
- We-pattern – expressions that use the first person plural nominative pronoun, such as 'We'll' or 'We're gonna'.
- Polite Directives – other expressions, such as 'Let's' or 'Let me'.

In this taxonomy, the first person singular pronoun patterns (I-pattern) represent the speakers' overt presence when expressing their communicative intentions. Polite directives and 'we-pattern' expressions would be considered more deferential alternatives that reflect the interactions between the lecturer and their student.

Ädel (2010)

Ädel has presented several considerable studies on metadiscourse in both written and spoken communications (Ädel, 2006, Ädel, 2010). The proposed scheme is the result of the analysis and combination of several existing schemes of metadiscourse, and is suitable for both written and spoken forms. Ädel's study was built using academic corpora that encompass both spoken and written discourse: the aforementioned MICASE corpus (Simpson and Swales, 2002), which consists of 30 spoken university lectures, from several disciplines, and 130 essays from the Michigan Corpus of Upper-level Student Papers (MICUSP) (Roḿer and Swales, 2009), written by highly proficient graduate students.

In contrast to previous studies on metadiscourse, Ädel (2010) focuses on highlighting the pragmatic function of the discourse, either in speech or in writing. In particular, the author seeks to find the rhetorical acts and recurrent linguistic patterns that speakers/writers often use to assist comprehension. That is, the study focused on comprehension to develop a scheme that represents the discourse function, rather than its form, and used these discourse functions as a guide to categorise metadiscourse. This general scheme can be applied to both varieties of discourse.

Table 3.1 shows Ädel's scheme of metadiscourse. It has four generic categories, namely Metalinguistic Comments, Discourse Organisation, Speech Acts, and Reference to the Audience. Each of these generic categories is further divided into several specific categories that

Metalinguistic Comments
Repairing
Reformulating
Commenting on Linguistic Form/Meaning
Clarifying
Managing Terminology
Discourse Organisation
Manage Topic
Introducing Topic
Delimiting Topic
Adding to Topic
Concluding Topic
Marking Aside
Manage Phorics
Enumerating
Endophoric Marking
Contextualising
Previewing
Reviewing
Speech Act
Arguing
Exemplifying
Others
Reference to the Audience
Managing Comprehension
Managing the Audience Discipline
Anticipating the Audience's Response
Managing the Message

TABLE 3.1: The metadiscourse scheme proposed by Ädel (2010).

signal their discourse functions. A detailed description of Ädel's scheme, with respect to these generic and specific categories, is provided below.

Metalinguistic Comments: includes five metadiscourse categories: *Repairing*, *Reformulating*, *Commenting on Linguistic Form/Meaning*, *Clarifying*, *Managing Terminology*. *Repairing* is used to correct or cancel a preceding contribution. Unsurprisingly, instances of this category can only be found in the MICASE spoken corpus, as demonstrated in the given examples, such as "I'm sorry". *Reformulating* is concerned with offering an alternative way to explain a previously stated idea to add value to it, not because the previous statement is wrong. This metadiscourse function was found in both spoken and written language. Examples include "In other words" and "Let me rephrase it a little". *Commenting on Linguistic Form/Meaning* provide comments on the linguistic form, word choice and/or meaning, such as "Can we put it in French language?", and can occur in both discourse modalities (spoken or written). This category is related to metasemantics, in particular the notion of 'mention' introduced by (Lyons, 1977). *Clarifying* is used to clarify the lecturer's intentions in order

to avoid any misunderstandings, and again is found in both discourse modalities. Examples include "I'm not saying that" and "For the sake of clarity I'm saying". Lastly, the function of *Managing Terminology* is used to give a definition for labels previously spoken about. This category is also related to 'mention' language (*e.g.* "We will be using the following definition").

Discourse Organisation: the functions under this generic label are further divided into two subcategories: *Manage Topic* and *Manage Phorics*. The discourse functions of the former share some similarities with the one proposed by Thompson (2003) described in Section 3.2.2. These are: *Introducing Topic*, *Delimiting Topic*, *Adding to Topic*, *Concluding Topic*, and *Marking Asides*. Lecturers often use the *Introduction* category to open new subtopics. For instance, in a physics lecture on Newton's Laws, "Newton's First Law", "Newton's Second Law" and "Newton's Third Law" would constitute subtopics. Conversely, the *Concluding Topic* category is normally used to conclude or summarise subtopics of the lecture, while the *Adding to Topic* category is used to add to the current subtopics, such as "I should add to that". *Delimiting Topic* expressions are used to establish a constraint when presenting the subtopic (*e.g.* "I am not going to cover this."). An example of this function would be "We're not going to deal with all eight here". Lastly, Marking Asides are used to open or close aside comments unrelated to either the topic or subtopics of the lecture, and only found in spoken discourse (*e.g.* "I want to do a little aside here.").

The other subcategory of **Discourse Organisation** *Manage Phorics*, consists of five specific functions, namely *Enumerating*, *Endophoric Marking*, *Previewing*, *Reviewing*, and *Contextualising*. *Enumerating* is used to show how specific parts of the discourse are ordered in relation to one another. An example of this category could be: "We're going to talk about mutations first", or "We want to deal with two things". *Endophoric Marking* shares some similarity with the Contextual category in Luukka's scheme, presented above, and is used to point to a specific location within the discourse; it refers to cases that have occurred before or after the present point in time (*e.g.* "Look at question number five in your handout"). This is in contrast to *Previewing* and *Reviewing*, which are used to point either forward or backward in the discourse. Examples include, "We'll be coming to that in the next lecture", or "we ended last time with". Finally, the *Contextualising* category is used to comment on the situation of speaking, and therefore contains traces of the production of the discourse itself (*e.g.* "We're doing pretty well on time".).

Speech Acts: these are composed of three specific metadiscourse functions: *Arguing*, *Exemplifying*, and *Other*. *Arguing* is used to express the action of arguing against some issue, such as "I argue that", while *Exemplifying* refers to situation when the lecturer explicitly demonstrated an example (e.g., "For example", or "I will use the example"). *Other*, as the name suggests, includes other discourse functions that are not observed frequently in

the dataset used in Ädel's study. Examples of this would be the categories suggesting, mentioning, and emphasising.

References to the Audience: this includes four metadiscourse functions related to interaction with the audience: *Managing Comprehension*, *Managing the Audience Discipline*, *Anticipating the Audience's Response*, and *Managing the Message*. Some of these functions can only found in spoken materials, such as *Managing Comprehension* and *Managing the Audience Discipline*. The former is used to ensure that both the lecturer and their audience understand one another; that is, it is used to refer to the audience's understanding regarding a particular point in the discourse, such as "Did I answer your question?", or "Can you guys hear?". The latter concerns cases where the lecturer wishes to have direct contact with the students to instruct them about something, such as complimenting them on their behaviour. An example of this would include "Can we have a little bit of quiet?" *Anticipating the Audience's Response* denotes strategies used by the speaker/writer to predict the audience's reaction regarding something that has been said, such as, "You might still think that", or "You guys will probably end up thinking". *Managing the Message* is often used to emphasise the core points, as well as to provide the 'big picture' regarding specific concepts within in the discourse. Examples of this function include, "What I want you to remember is", or "The take-away message is". Finally, Imagining Scenario is used to engage the students with a particular experiment that requires mutual thought between the lecturer and the student such as "Can you imagine?", "So, suppose ...", or "Imagine the following scenario".

It is important to note that the scheme by Ädel (2010) is the only metadiscourse scheme that is designed to present function-oriented categories at sentence-level, which is the main interest in this thesis, as discussed in Chapter 1. In addition, it has proved to be an effective scheme in annotating metadiscourse in spoken language using presentation-style datasets such as TED Talks, as reported by Correia et al. (2014b, 2016), and further discussed in Section 2.2.1. In fact, the work presented by Correia et al. (2016) is the study most closely related to the work conducted in this thesis. However, the target dataset is different, as this study examines the phenomena in terms of university lecture courses that contain a set of related lectures and topics from the OCW platforms. This type of data fits better with the scheme proposed by Ädel (2010), as it was also the target dataset in the small analytical experiment in that study, as discussed above.

3.3 Annotation Method

The previous section has shown that Ädel's scheme describes a functional approach to metadiscourse that is suitable for analysing the discourse of academic lectures from different disciplines. For this reason, this thesis has adopted this scheme when building a corpus of

metadiscourse for academic lectures. However, building a corpus is a task that consists of several sequential steps in the annotation task. First, one needs to select the annotation scheme that matches the motivations for building the corpus. For example, the main intuition behind building a corpus of metadiscourse in academic lectures was due to the absence of a suitable lecture resource that could be used to develop the metadiscourse tagging approach; this motivated the decision to build a corpus specifically for OCW academic lectures from different disciplines.

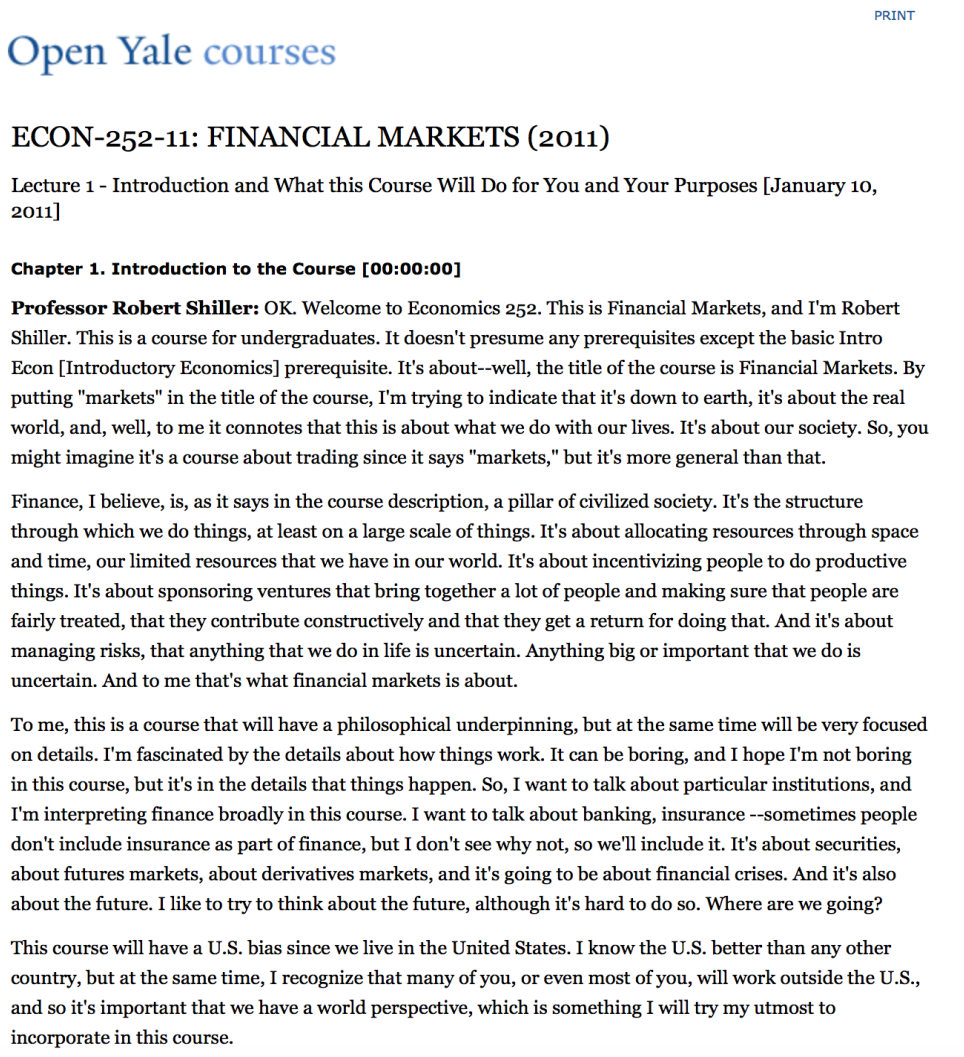
The next step is to select and prepare the lecture datasets to be annotated with the categories of the chosen scheme – that is, the OCW academic lectures described in Section 2.4. Then, the participants who will perform the annotation task must be selected. In this annotation experiment we recruited expert annotators. Finally, a preliminary annotation experiment must be conducted to check whether the annotators understand the task clearly, and whether the categories of the selected scheme fit the chosen datasets. The following sections describe the process for each of these steps for building a corpus of metadiscourse for academic lectures.

3.3.1 Scheme

As mentioned above, the scheme used is the one proposed by Ädel (2010), described fully in Section 3.2.2. This is because Ädel's scheme combines several previous approaches to metadiscourse and meet the following criteria:

1. Its categories signal high-level discourse functions, which is beneficial for discourse analysis related tasks such as the thematic discourse function presented in this thesis.
2. It is suitable for spoken discourse, in particular academic lectures across disciplines, as MICASE lectures were used in developing this scheme.

Previous work has adopted this scheme to build a corpus of metadiscourse using TED Talks (Correia et al., 2014b). In their study Correia et al. (2014b) combined some categories that serve the same metadiscourse functions, such as the integration of the specific category Exemplifying, in the Speech Acts Labels with the category Imagining Scenarios, in the Interaction with Audience general category. A further integration was also made in the present study of the categories Emphasising in the **Speech Acts** labels, and Managing the Message, in Interaction with Audience. This is because these two categories convey quite similar metadiscourse functions. In addition, it is hypothesised that the category Suggesting, in **Speech Acts** labels, will consistently occur in academic lectures. These decisions were based on several training trials with the annotators, as similarities were noted between these



PRINT

Open Yale courses

ECON-252-11: FINANCIAL MARKETS (2011)

Lecture 1 - Introduction and What this Course Will Do for You and Your Purposes [January 10, 2011]

Chapter 1. Introduction to the Course [00:00:00]

Professor Robert Shiller: OK. Welcome to Economics 252. This is Financial Markets, and I'm Robert Shiller. This is a course for undergraduates. It doesn't presume any prerequisites except the basic Intro Econ [Introductory Economics] prerequisite. It's about--well, the title of the course is Financial Markets. By putting "markets" in the title of the course, I'm trying to indicate that it's down to earth, it's about the real world, and, well, to me it connotes that this is about what we do with our lives. It's about our society. So, you might imagine it's a course about trading since it says "markets," but it's more general than that.

Finance, I believe, is, as it says in the course description, a pillar of civilized society. It's the structure through which we do things, at least on a large scale of things. It's about allocating resources through space and time, our limited resources that we have in our world. It's about incentivizing people to do productive things. It's about sponsoring ventures that bring together a lot of people and making sure that people are fairly treated, that they contribute constructively and that they get a return for doing that. And it's about managing risks, that anything that we do in life is uncertain. Anything big or important that we do is uncertain. And to me that's what financial markets is about.

To me, this is a course that will have a philosophical underpinning, but at the same time will be very focused on details. I'm fascinated by the details about how things work. It can be boring, and I hope I'm not boring in this course, but it's in the details that things happen. So, I want to talk about particular institutions, and I'm interpreting finance broadly in this course. I want to talk about banking, insurance --sometimes people don't include insurance as part of finance, but I don't see why not, so we'll include it. It's about securities, about futures markets, about derivatives markets, and it's going to be about financial crises. And it's also about the future. I like to try to think about the future, although it's hard to do so. Where are we going?

This course will have a U.S. bias since we live in the United States. I know the U.S. better than any other country, but at the same time, I recognize that many of you, or even most of you, will work outside the U.S., and so it's important that we have a world perspective, which is something I will try my utmost to incorporate in this course.

FIGURE 3.1: Example of raw transcripts of an Economic lecture, provided by Shiller (2011).

categories and this caused some confusion for the annotators. The final pilot study therefore had a total of 22 metadiscourse functions, as described below.

3.3.2 Datasets

Having deciding on the metadiscourse scheme to use in the annotation experiment, the next step is to decide on the source of academic lectures from different disciplines. As part of the objective of this thesis is to facilitate access to the OCW materials online by developing a tool for discourse analysis, such as for the metadiscourse tagging task, the output of this task will be used in downstream applications such as thematic discourse segmentation, as described in Chapter 7. For these reason, the datasets used in this annotation experiment are OCW lecture courses in Physics and Economics. Further details about this collection of OCW datasets, alongside the main objectives of this theses, are described in Section 2.4.

However, there are some pre-processing step that must be applied to the datasets prior to the annotation process. For instance, the reference transcripts of the datasets are written at paragraph level, wherein each paragraph may represent a point in the discussion. Figure 3.1 shows examples of the raw lecture transcripts from the Economics courses. These transcripts had to be split at sentence level as part of preparing the data for the annotation task, using the Stanford Tokeniser[1].

3.3.3 Participants

After deciding on both the scheme and the dataset to use, the next step is select the annotators. Four expert on-site participants with domain-specific knowledge took part in this study, along with the author of this thesis, who is familiar with metadiscourse phenomena. This contrasts with previous work on metadiscourse using TED Talks, which instead hired non-expert annotators via a crowdsourcing service. This is because the content of academic lectures is more complex than general talks and needs some previous knowledge of the materials presented in order to annotate instances of metadiscourse accurately. The invited annotators were students, two working towards a PhD in Physics and the other two towards a PhD in Economics. Thus, the annotators were mostly familiar with the introductory subject matter of the discipline lectures. During the preparatory stage of the experiment, the annotators familiarised themselves with the annotation scheme, which included various examples of every metadiscourse category.

3.3.4 Pilot Study

Once the scheme, source and participants have been established, the next step is to conduct a trial study to investigate the appropriateness of the scheme when applied to the dataset. The aim of this small annotation study is to gain an overview of the density of metadiscourse categories in the given dataset. The final set of categories to be used to build the corpus can then be decided, based on this trial study. In this initial study, five lectures were selected at random from each discipline, to be annotated with Ädel's categories, as described in Section 3.2.2. All the participants took part in this initial study in order to train them for the final task. The list of occurrences was finalised according to overall agreement, as will be described in Section 3.3.7. Below, a description of the distribution of the metadiscourse categories, as shown in Table 3.2, along with some examples in the sample dataset, is provided. This discussion was organised based on the four generic categories in Ädel's scheme.

[1] http://nlp.stanford.edu/software/tokenizer.shtml

		Physics	Economics
	Category	**Occurrences**	**Occurrences**
Metalinguistic	Repairing	9	11
	Reformulating	16	13
	Commenting on Linguistic Form/Meaning	8	7
	Clarifying	19	17
	Managing Terminology	25	18
	Total	77	66
Discourse Organisation	Introduction	25	54
	Conclusion	20	19
	Adding to Topic	2	5
	Delimiting	5	21
	Contextualising	3	2
	Marking Aside	-	1
	Enumerating	18	21
	Endophoric	2	6
	Reviewing	50	33
	Previewing	68	44
	Total	193	206
Speech Acts	Emphasising	113	98
	Exemplifying	88	95
	Arguing	9	4
	Suggesting	1	3
	Total	211	200
Audience	Managing Comprehension	21	14
	Managing Audience Discipline	2	-
	Anticipating Audience's Response	11	8
	Total	34	22
	Overall	515	494

TABLE 3.2: Number of occurrences organised by discipline for each metadiscourse category in the pilot study.

Metalinguistic Comments

Metalinguistic Comments are compose of five specific metadiscourse functions, mainly used to comment on the use of the language – either its form or meaning. All of the five categories occurred frequently, particularly the category *Managing Terminology* as indicated in Table 3.2. The instances of these functions in the sample indicate that lectures may require less preparation when compared with a presentation-based task, such as TED Talk, where these functions rarely occur. Some examples of these functions from the sample are illustrated below.

- *Repairing*

  - **I'm sorry that should have been** – good catch.

- *Reformulating*
  - **In other words**, you take all the profits.
- *Commenting on Linguistic Form/Meaning*
  - **I'll try to say it in** Chinese.
- *Clarifying*
  - **I'm not saying you have to** have strong preferences.
- *Managing Terminology*
  - **But what we mean by** short run here is no firm entry and exit.

Discourse Organisation

From the metadiscourse functions of the first subcategory, *Manage Topic*, only *Introducing Topic* and *Concluding Topic* were found frequently in both disciplines. However, *Delimiting* was found more consistently in the Economics sample than in Physics. The function *Marking Aside* was not found at all in the Physics lectures and only once in Economics. This may reflect that these materials are lengthy lectures, with an average of 50 minutes; they therefore contain multiple sub-topics that need strategies to organise them, to allow the students to digest the information given. Overall, it appears that Economics lecturers use discourse functions more frequently to organise their lectures than those in Physics.

Regarding the other functions of the subcategory *Manage Phorics* in *Discourse Organisation*, significant occurrences of the functions *Enumerating*, *Previewing*, and *Reviewing* were found in both disciplines, compared to *Contextualising*. This can be attributed to the fact that in lecture courses lecturers may review some points from the previous lecture or preview what is coming in the next. The *Endophoric* function seems to occur slightly more often in Economics than in Physics, as the Economics lecturers may refer to charts and figures in their slides, while the Physics lecturer use the blackboard instead. Examples of *Discourse Organisation* functions represented in the lectures sample (from both Physics and Economics) are:

- *Introducing Topic*
  - **Now I wanted to move to the third topic ,which is** state and local finance
- *Delimiting Topic*

  - **We're not going to cover** model labour .

- *Concluding Topic*
  - **Let me just conclude with this** part.
- *Enumerating*
  - **The first one is** just about the morality of finance .
- *Previewing*
  - **And in the next lecture we're going to** use this again
  - **I'm coming to that in a minute** .
- *Reviewing*
  - **Last lecture I introduced the concept of** angular momentum and torque .
  - **I told you in the beginning of the lecture that** what is going to happen.

Speech Act Labels

As discussed in Section 3.2.2, this category consists of three discourse functions: *Arguing*, *Exemplifying*, and *Other*. From the *Others*, *Emphasising* and *Suggestion*, were added to the list. Both *Exemplifying* and *Emphasising* have significant representation in the sample in both disciplines. *Arguing* and *Suggestion* occur less frequently in the sample, in particular Suggestion in Physics lectures.

These functions were found consistently throughout the ten lecture sample.

- *Arguing*
  - **I want to argue** without writing any questions that if energy and momentum are conserved.
- *Suggesting*
  - **I would suggest that** the Barron's articles really took far too short a time horizon .
- *Exemplifying*
  - **For example**, when you want to buy eggs you measure in dozens.
  - **So imagine** a tiny segment of width.

	Category	Abbreviation
Metalinguistic	Repairing	REP
	Reformulating	REF
	Commenting on Linguistic Form/Meaning	CLF
	Clarifying	CLA
	Managing Terminology	MAT
Discourse Organisation	Introduction	INT
	Conclusion	CON
	Delimiting	DEL
	Contextualising	COT
	Enumerating	ENU
	Endophoric	PHO
	Reviewing	REV
	Previewing	PRE
Speech Acts	Emphasising	EMP
	Exemplifying	EXE
	Arguing	ARG
	Suggesting	SUG
Audience	Managing Comprehension	MAC
	Anticipating Audience's Response	AAR

TABLE 3.3: The final set of the metadiscourse categories used in this thesis, adapted from Ädel (2010), organised based on metadiscourse generic labels, along with abbreviations for each category.

- *Emphasising*
  - **Now a very important point to notice is that** this whole thing ...
  - **You should remember** this equation .

Reference to Audience

The nature of academic lectures requires the lecturer to engage with their students, for example answering questions or making sure that they are abreast of the concepts introduced in the lecture. For this reason, metadiscourse functions related to Reference to Audience or Interaction with the Audience (herein) are found consistently in the sample across disciplines, in particular, the functions *Managing Comprehension* and *Anticipating Audience's Response*. On the other hand, functions that discipline the students, such as *Managing Audience Discipline*, were rarely found to occur in either discipline. Examples are:

- *Anticipating Response*

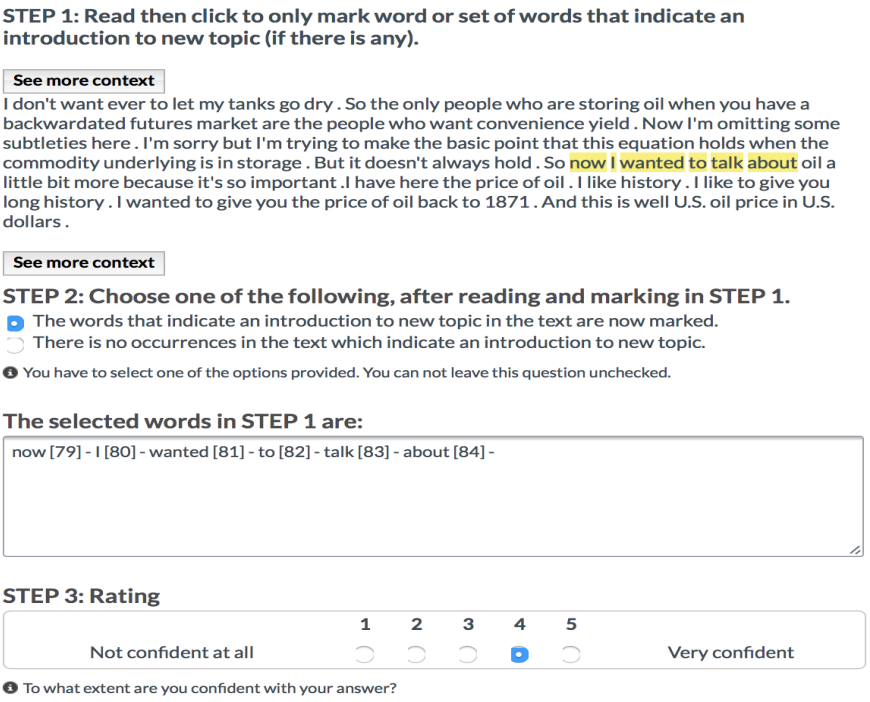


FIGURE 3.2: Example of the annotation interface used in annotating the metadiscourse category *Introduction*

  - **You might think they have** a trajectory but they don't .

- *Managing Comprehension*
  - **Can you hear that ?**

From the outcome of this trial study, two criteria formulate the final set of metadiscourse functions in the annotation experiment: 1. the most frequent functions in the trial study, and either common in both disciplines or consistently occurring in one of the disciplines; 2. the input from the literature and spoken discourse analysis studies in general. On this basis, the discourse functions included in the final set of annotations for building the corpus are 19 in total. These are shown in Table 3.3 alongside the abbreviations for these functions that will be used from here on in this thesis. As before, the generic category labels are retained and the specific metadiscourse functions are organised according to this generic set. These generic categories are: *Metalinguistics*, *Discourse Organisation*, *Speech Acts*, and *Interaction with Audience*. This is because later in the modelling stages (Chapters 5 and 6) we investigate how the model performs at two levels of granularity: generic and specific metadiscourse functions.

3.3.5 Tools and Guidelines

In order to facilitate the process for the annotators, the 19 metadiscourse categories are annotated, one at a time, with only one segment with an average of 200 words per task (truncated to the closest end of utterance). Thus, for every category in the annotation scheme, there are a total of 2,440 annotation tasks for the Physics lectures and 2,110 annotation tasks for the Economics lectures. The purpose of creating these large numbers of annotation tasks for each category is to simplify the process for the annotators and hence reduce the cognitive loads on them when performing the tasks. In addition, there are different instruction sets prepared for each of the 19 metadiscourse categories in the scheme, along with examples. These instructions are given to the annotators before starting the annotation task; an example of these sets of instructions is provided in Figure A.1 in Appendix A, for the metadiscourse category Emphasising. The purpose of including a set of correct and incorrect examples along with the instructions is to increase the level of understanding of the task among annotators.

The annotation task is conducted with the help of an online annotation tool, which is also useful in outlining further brief instructions with each task. The online tool was created and designed specifically for this task using HTML/XML languages and JavaScript functions. Moreover, specific mechanisms were provided in order to facilitate the work of the annotators, such as requesting them to highlight the target word or set of words that they consider are indicators of the desired metadiscourse category. The annotation interface for the category Introduction is demonstrated in Figure 3.2, which shows how the instructions were integrated with the annotation tool and consist of three essential steps: first reading the text segment and highlighting the metadiscourse tag occurrences, then confirming this finding in the second step, and finally asking the annotator to rate their confidence score regarding this particular metadiscourse tag. It is worth noting that the final set of instructions were derived after several preliminary trials.

3.3.6 Agreement Measure

The agreement measure selected is the one most commonly applied in NLP research, *Kappa* (Carletta, 1996), specifically Fleiss' kappa coefficient κ Fleiss (1971).

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

where $1 - \bar{P}_e$ provides agreement that is attainable above chance, while $\bar{P} - \bar{P}_e$ provides agreement that actually achieved above chance. In this context, complete agreement corresponds to $\kappa = 1$, and no agreement corresponds to $\kappa \leq 0$.

category	MD	Transcript
	REV	**Let me remind you that everything I did so far** in class came from analyzing the Lorentz transformations.
	EMP	**And you guys should be really** on top of those two marvelous equations because all the stuff we are doing is a consequence of that.
	DEL|REV	**I won't go over how** we derive them because **I've done it** more than once.
	REV	**But I remind you that** if you've got an event that occurs at x_t for one person and to a person moving to the right at velocity the same event will have coordinates $\grave{x} = x - u_t$.
	NONE	This is it .
	EMP	**This is the key** .
	NONE	From this by taking differences of two events you can get similar equations for coordinate differences.
	REF	**In other words** if two events are separated in space by x for one person and $\grave{x}$ for another person and likewise in time then you get similar formula for differences.
	NONE	So differences are related the same way that coordinates themselves are.
	NONE	But I will write it anyway because I will use it sometimes one way and sometimes the other way.
	NONE	Even this one you can think of as a formula for a difference except one of the coordinates or the origin.
	NONE	So I put this to work.
	NONE	I got a lot of consequences from that.
	REV	**I You remember that**
	EXE|REV	**For example** **I said** take a clock that you are carrying with you.
	NONE	Or let's take a clock that I'm carrying with me and let's see how it looks to you.
	EXE	**And let's say** it goes tick and it goes tick one more time.

TABLE 3.4: Example from Physics lecture yale-phy0020014. 'MD' denotes metadiscourse category (multiple tags are separated by '|'). For the purposes of illustration, each category in the table is indicated by a unique font colour, and the phrases that reflect that have been highlighted with the same colour. The author of the thesis annotated these expressions.

3.3.7 Gold Standard

Since this study aims to detect whether an utterance contains an instance of metadiscourse function, the decision as to whether or not an utterance contains the discourse function of a certain metadiscourse category is made based on overall agreement between annotators. More precisely, the agreement on the final annotation (considering all utterances) between the three annotators (the thesis author and the other two domain specialists as mentioned in Section 3.3.3) were based on two specific steps: 1. check if two annotators agree that this particular utterance contains a particular metadiscourse tag, based on majority vote, then 2. check if the intersection (in terms of the number of words on the same utterance) between their annotations is not void. For example, two annotators still agree when one selects "Today, the topic of the lecture is" and the other misses some of the words, selecting "the topic of the lecture is". A stricter approach for computing the agreement at word-level is reported in Madnani et al. (2012) for an extraction task of metadiscourse expressions,

category	MD	Transcript
	INT	**So let's talk about** our preference assumptions.
	ENU	So to model consumers preferences across goods we're going to impose **three** preference assumptions.
	ENU	**There are three** preference assumptions.
	ENU|REV	**Assumption one** now once again **let me remind you from the first lecture** this is getting to some of the harder material.
	NONE	I'm going to write messily and talk quickly so stop me if anything is.
	NONE	And if you don't stop me I'll just go faster and faster until I explode.
	NONE	So basically feel free to interrupt and stop me with questions and such.
	ENU	**There are three** assumptions on preferences.
	ENU	**The first assumption is** the assumption of completeness.
	NONE	When comparing two bundles of goods you prefer one or the other but you don't value them equally.
	NONE	OK when comparing two bundles of goods you prefer one or you prefer the other but you're not indifferent.
	NONE	Completeness is the same as no indifference.
	CLA	**So what we're saying is** whenever I offer you two bundles of goods you could always tell me what you like better.
	NONE	Now it could be infinitesimally better.
	CLA	**I'm not saying** you have to have strong preferences.
	NONE	There always at least some slight preference for one bundle of goods over another.
	NONE	That's the completeness assumption.
	NONE	Now in reality often times we are indifferent.
	NONE	Well once again his is one of these simplifying assumptions that will make the model work.
	NONE	We're just going to say more precisely you are never purely indifferent.
	REP	**I'm sorry** let me back up.
	PRE	Forget I said indifferent because we'll want to use that word in a different context **later in the lecture** .

TABLE 3.5: Example from Economics lecture mit-eco0020023. 'MD' denotes metadiscourse category (multiple tags are separated by '|'). For the purposes of illustration, each category in the table is indicated by a unique font colour, and the phrases that reflect that have been highlighted with the same colour. The author of this thesis annotated these expressions.

not classifying them from the sentence. This is different from the metadiscourse tagging task presented in this work as the main intuition here is to assign tags to instances of the metadiscourse in lecture discourse. It is important to note that on this final run of the annotation experiment where the gold standard is formulated, the annotators, especially the domain specialist, become more familiar with the task and gain more experience on the task compared to the previous trial studies as discussed in Section 3.2. This in turn increases the level of agreements between annotators, as will be shown in the next section.

Examples of the gold standard set for both Physics and Economics are provided in Table 3.4 and Table 3.5, respectively. These examples serve as an illustration of some of the types of metadiscourse that one is able to observe in the generated corpus. The example also provides a taste of the frequency and complexity of metadiscourse information. For instance, in Table 3.5 , of the 17 utterances within the excerpt, 8 have no labels; this is typical of many regions

		Physics		Economics		Average	
	MD Tag	κ	**Confidence**	κ	**Confidence**	κ	**Confidence**
Metalinguistic	REP	0.76	3.80	0.73	3.60	0.75	3.70
	REF	0.83	3.91	0.75	3.68	0.79	3.79
	CLF	0.70	3.74	0.73	3.80	0.72	3.77
	CLA	0.69	3.56	0.66	3.35	0.68	3.46
	MAT	0.77	3.86	0.79	3.82	0.78	3.84
	Total	0.75	3.75	0.73	3.65	0.74	3.71
Discourse Organisation	INT	0.85	3.98	0.80	4.00	0.83	3.99
	CON	0.79	4.00	0.77	3.80	0.78	3.90
	DEL	0.80	3.93	0.74	3.76	0.77	3.85
	COT	0.70	3.86	0.71	3.63	0.71	3.75
	ENU	0.78	3.76	0.81	3.89	0.79	3.83
	PHO	0.75	3.89	0.78	3.78	0.77	3.84
	REV	0.80	3.97	0.81	3.98	0.81	3.98
	PRE	0.77	3.81	0.76	3.85	0.77	3.83
	Total	0.78	3.90	0.77	3.84	0.78	3.87
Speech Acts	EMP	0.81	4.13	0.84	4.20	0.83	4.17
	EXE	0.82	3.97	0.85	4.00	0.84	3.99
	ARG	0.75	3.89	0.67	3.55	0.71	3.72
	SUG	0.74	3.83	0.72	3.69	0.73	3.76
	Total	0.78	3.95	0.76	3.86	0.78	3.91
Audience	MAC	0.71	3.78	0.69	3.71	0.70	3.75
	AAR	0.80	3.99	0.74	3.90	0.77	3.95
	Total	0.76	3.89	0.72	3.80	0.74	3.85
	Average	0.77	3.87	0.75	3.79	0.76	3.83

TABLE 3.6: Results organised based on discipline, in terms of inter-annotator agreement (Fleiss' kappa κ) and the self-reported confidence scores for each metadiscourse category. The average row-wise refers to scores across disciplines, while the average column-wise represents scores across categories.

in the corpus, which show no metadiscourse tags at all. In light of this, NONE was by far the most common label. Further, some sentences have more than one metadiscourse category (referred to as multifunctions from herein) such as the case with the third utterance, which has two tags: *Delimiting* (DEL) and *Reviewing* (REV).

3.4 Annotation Results and Analysis

This section presents the results of the final annotation experiment for each of the metadiscourse categories considered in this task. Tables 3.6 and 3.7 summarise the results for each metadiscourse category. Table 3.6 shows the inter-annotator agreements, along with behavioural data such as the self-confidence score (averaged on a 5-point Likert scale) for Physics and Economics lectures. Table 3.7, meanwhile, shows the number of occurrences for

each metadiscourse category in the obtained gold standard sets for both disciplines. Note that the number of occurrences is the number of sentences that were annotated with the metadiscourse category.

3.4.1 Inter-annotator Reliability

The seventh column of Table 3.6 shows the cross-discipline agreement for each metadiscourse category. In general, the inter-annotator agreements scores are in the range $\kappa \in [0.68 - 0.84]$, which indicates high agreement. Those with > 0.80 such as INT, REV, EMP and EXE are considered to show substantial agreement. This may be due to the fact that annotators dealt with each category one at a time, so the probability of two annotators selecting the same sentence by chance is very low. However, categories such as CLA show low agreement between annotators compared to others categories, consistent across both disciplines. Looking at the data, the most common structure for this category has the form of: "I am sorry ... what I mean ...", so it is possible that some annotators did not considered this as instances of CLA, more likely labelling it as REP.

It is also interesting to note that when analysing the results based on the four generic categories, the metadiscourse categories of **Discourse Organisation** and **Speech Acts** reported the best inter-annotator agreement, with $\kappa = 78$ across disciplines. This may reflect the number of occurrences, as will be seen next, since these two generic categories contain most of the labelled sentences. Another possible reason is that the metadiscourse patterns of these types of categories are very clear and standard, compared to either the **Metalinguistics Comments** or **Reference to Audience**, as demonstrated in Table 3.4 and Table 3.5, for Physics and Economics lectures, respectively. In sum, the differences between these sets of categories can, in fact, be subtle, which may justify the low-level agreement. On the other hand, the categories of **MetaLinguistics Comments** required some thought to differentiate between them which hindered the annotation process little bit. Similar observations were made with the two categories of **Reference to Audience**, namely MAC and AAR, as both required some expression patterns that involve interaction with students, which are sometimes mixed in one sentence. In other words, the sentence can hold two categories not because it has two metadiscourse expressions, but because it has one expression that can cover two categories. For example, "You guys will probably be asking if we understand this." Although such cases are infrequent in our dataset, as indicated by the overall agreement for these categories, they can affect the agreement scores, as in MAC, in particular in Economics lectures, with $\kappa = 0.69$.

	MD category	Physics		Economics		Overall	
		#	%	#	%	#	%
Metalinguistic	REP	83	1.36	94	1.72	177	1.54
	REF	181	2.97	71	1.30	252	2.19
	CLF	18	0.29	37	0.68	55	0.48
	CLA	285	4.68	300	5.50	585	5.10
	MAT	545	8.95	319	5.85	864	7.54
	Total	1112	20.96	821	15.05	1933	16.85
Discourse Organisation	INT	208	3.42	346	6.35	554	4.83
	CON	104	1.71	123	2.26	227	1.98
	DEL	87	1.43	82	1.50	169	1.47
	COT	22	0.36	31	0.57	53	0.47
	ENU	571	9.38	583	10.70	1154	10.06
	PHO	123	2.02	195	3.58	318	2.77
	REV	834	13.70	685	12.57	1519	13.25
	PRE	536	8.81	396	7.27	932	8.13
	Total	2485	40.83	2441	44.80	4926	42.92
Speech Acts	EMP	1234	20.27	1070	19.63	2304	20.09
	EXE	842	13.83	885	16.24	1727	15.06
	ARG	43	0.71	14	0.26	57	0.49
	SUG	11	0.20	22	0.40	33	0.29
	Total	2130	35.01	1991	36.53	4121	35.94
Audience	MAC	218	3.58	110	2.02	328	2.86
	AAR	113	1.86	45	0.83	158	1.38
	Total	331	5.44	155	2.85	486	4.24
	Overall	6058	18.93	5408	17.45	11466	18.20

TABLE 3.7: A statistical summary of all the categories in the gold standard dataset for each discipline, showing the number of occurrences (#) and the frequency of each category relative to all other categories (%). The overall row-wise is the scores across disciplines, while the overall column-wise represents scores across categories.

3.4.2 Self-reported Confidence Score

The next results to analyse are the self-reported confidence scores, as presented in Table 3.6, in particular in columns 4 and 6 for Physics and Economics, respectively. The last column shows the average scores across both disciplines on a 5-point Likert scale. All metadiscourse categories score above the middle of the scale (3), with annotators being less confident for categories such as CLA in both disciplines. Interestingly, CLA tag shows lower agreement scores as well, as indicated above. From this its seems there is a relationship between having high confidence scores and high agreement. Another interesting example of such correlation is with the categories INT, EMP, and EXE, as these three categories scored the best for agreement, and also have the highest self-reported confidence scores. This indicates that the annotators understand the task of annotating metadiscourse instances for such categories. This observation is also noticed at the generic level; for example, on average both

Discourse Organisation and **Speech Acts** have high agreement and high self-confidence scores. Conversely, annotators show low confidence for both **MetaLinguistics Comments** and **Reference to Audience**. Again, the results of these generic labels give an indication that these categories may need a wider context to identify them.

3.4.3 Metadiscourse Occurrences

Table 3.7 represents the number of sentences labelled with metadiscourse categories. This information is again organised according to the generic categories for each disciplines, then overall. Among all categories, the most frequently occurring are those belonging to both **Discourse Organisation** and **Speech Acts**. Notably, these categories also generally have significant agreement and the highest self-reported scores across disciplines, as noted above. **Discourse Organisation** and **Speech Acts** have in total 4,926 and 4,121 occurrences, respectively. However, this is not the case at the specific level of metadiscourse categories. For example, the category PRE in both disciplines has a high number of occurrences (932) but a lower agreement score ($\kappa = 0.77$) when compare to CON, which has a slightly higher agreement score of $\kappa = 0.78$ but a far lower number of occurrences (227). In other cases, categories have similar agreement scores but vary in the number of occurrences, such as the category ENU compared to REF. In sum, the relationship between the inter-annotator agreement and self-reported confidence scores and the number of occurrence is not inclusive. This is because there are some other factors that may have impact on these results, such as the clarity of the category itself, since some categories are more confusing than others.

Other important factors are the effect of domain knowledge and lecturers' styles. For instance, although the number of lectures in the Physics domain (57) is more than those for Economics (only 49), the category INT (an act used normally to introduce a new topic) occurred more frequently in Economics than in Physics. This can be attributed to the fact that lecturers in social science may tend to structure their lecture contents into multiple subtopics, in order to allow students to digest the given information. However, all in all, the number of lectures does not appear to correlate with the number of introduced topics, as lectures may have several subtopics. Moreover, such observations also align with what was encountered in the trial study, shown in Table 3.2. In addition, the total number of occurrences in the pilot study for metadiscourse categories of both **Discourse Organisation** and **Speech Acts** are higher than those for **MetaLinguistics Comments** and **Reference to Audience**.

3.4.4 Discussion

To date, several corpora have been proposed to fills the gap in discourse analysis for the research community. However, only a few have addressed the task for spoken discourse – or, more precisely, provided a reference of the strategies used by the speakers to structure their discourse content. Therefore, the aim of the OpenCourseWare metadiscourse (OCWMD) corpus is to contribute to such analysis by providing a case study for academic lectures. It is composed of a set of 19 metadiscourse categories used to label sentences by expert annotators. Experts were hired because the subjects of the lecture materials necessitated a level of domain knowledge to pursue the annotation task. The informed annotators showed the ability to annotate metadiscourse phenomena in academic lectures. In part this was assisted by the task design, which meant that annotators were asked to only process one metadiscourse category at a time, for a segment of only 200 words, with clear instructions – as demonstrated in Section 3.3.

Overall, the annotation experiment shows that the level of agreement between participants is higher ($\kappa \in [0.68 - 0.84]$) than for a previous study on the same lectures dataset, but with fewer categories ($\kappa \in [0.60 - 0.71]$), as discussed in Section 3.4.1. This indicates that the annotators' knowledge about the task may have increased. However, there is still a small amount of disagreement among annotators regarding the category CLA in both disciplines. Previous related annotation studies, such as by Correia et al. (2015, 2016) working on TED Talks, show similar observations with the CLA category. In the second of these studies this was attributed to the span of occurrences: as two statements are part of the same instance of a CLA, they can be split into two segments when annotated by non-expert annotators. In this study it seems more likely to be due to the fact that the annotators have confused this category with REP, as sometimes both occurred within one sentence. According to Landis and Koch (1977), agreement can be represented on the scale as ($\leq$ 0 no agreement– [0-0.20] slight – [0.21-0.40] fair – [0.41-0.60] moderate – [0.61-0.80] substantial – [0.81-1] almost perfect). Based on this metric, the agreement scores obtained in our datasets can be considered substantial for most of the categories.

The occurrences of metadiscourse tags have been reported in terms of absolute numbers and relative frequencies for all the tags in the scheme. In total, the annotated corpus has 11,466 occurrences of metadiscourse across all categories and disciplines. Physics has a slightly higher number of occurrences (6,058) than Economics (5,408). The impact of domain knowledge can be seen when analysing the agreement and self-confidence scores. For instance, in Physics lectures, on average annotators gave a higher agreements score of $\kappa = 0.77$ and a confidence score of 3.87, compared to $\kappa = 0.75$ and a confidence score of 3.79 in Economics. Another interesting observation from the results is that most of the utterances in the corpus have been labelled with NONE. This is due to the fact that these metadiscourse tags form

the functional structure of the discourse, and therefore serve as brackets for the propositional content of the discourse (Schiffrin, 1980). This case seems typical in comparison with related work, for example in identifying metadiscourse in a presentation-style corpus (Correia et al., 2014b), or detecting speech acts in messages (Qadir and Riloff, 2011).

3.5 Conclusion

This chapter presents the first stage of the metadiscourse tagging approach described in the previous chapter. This stage is about building a OCWMD corpus of metadiscourse in academic lectures from two different disciplines: Physics and Economics. Metadiscourse phenomena are commonly defined as linguistic expressions that often refer to discourse about discourse, and signal the function of the discourse. Previous attempts to build a corpus of metadiscourse for spoken discourse used TED Talks. This study has employed an existing scheme designed to provide a function-oriented taxonomy of metadiscourse. This existing scheme merged previous approaches to metadiscourse taxonomies, to be applicable to both spoken and written discourse.

An investigation of the use of this scheme on academic lectures from the two disciplines has been described. The objective has been to study the effects of discipline knowledge on the annotation task. The use of these tags or categories as features for recovering the higher level structure of lectures will be explored, such as the task of thematic discourse segmentation, which will be demonstrated in Chapter 7. Nevertheless, a further adaptation process on this scheme was applied, based on a trial study. This adaptation takes into account both the material to annotate and the setting in which the annotation task is performed. Experiments with the selected OCW lecture datasets described in the previous chapter show that expert annotators are able to identify occurrences of multiple categories of metadiscourse, and hence confirm a reliable coding of metadiscourse in academic lectures using the adapted annotation scheme.

Chapter 4

Automatic Transcriptions of Academic Lectures

The previous chapter presented the first stage of the metadiscourse tagging approach developed throughout this thesis. A corpus of metadiscourse has been built using reference transcriptions to be used in the development of the metadiscourse tagging model. The aim was to investigate the occurrences of metadiscourse phenomena within academic lectures in two disciplines: Physics and Economics. This chapter describes how to produce automatic transcriptions of academic lectures by building an automatic speech recognition (ASR) system specifically for OCW lectures resources. This ASR-OCW system is part of the metadiscourse tagging approach aiming to automate all of its aspects, and has two components, the acoustic model (AM) and language model (LM). The AM was trained on a set of 421 hours of academic lectures speech. The focus is on the LM adaptation by the integration of many in-domain and out-of-domain resources. In order to appropriately evaluate the ASR-OCW system, a manual reference with word timings was required. For that purpose, a lightly supervised alignment approach was used. Experiment results on the Physics and Economics lectures set indicate that the ASR system with interpolated LM achieved a Word Error Rate (WER) of 28

The chapter is structured as follows: Section 4.1 introduces the ASR-OCW system for lectures, and the lightly supervised alignment with the reference transcriptions as the task used to evaluate this stage, along with the motivations for this approach and its contribution. Section 4.2 reviews previous work related to language model adaptation techniques and previous language model adaptation approaches for academic lectures. The implementation of the proposed ASR-OCW system and lightly supervised systems is presented in Section 4.3. A summary of the results obtained from the evaluation experiment is given in Section 4.4 Section 4.5 provides a conclusion for the chapter.

4.1 Introduction

The processing of academic lecture recordings has become an area of interest for both research and application communities. Improving access to recorded lecture archives is a task that involves research efforts common to both Automatic Speech Recognition (ASR) and Human-Computer Interaction (HCI). Such technologies support applications that further enhance the accessibility of these recordings. The main intuition is driven by the use of automatic transcripts in downstream applications, such as efficient browsing by indexing video with audio transcripts. This is in addition to other speech-enabled applications, such as information retrieval and summarisation, or translation of the lecture content. Moreover, they allow for far more complex tasks, such as the development of metadiscourse systems that require a level of spoken discourse understanding.

Nowadays, several projects initiated by non-profit organisations aim to support e-learning applications. They provide online platforms to freely access lecture materials, such as OCW project. The OCW project attracts many universities from across the globe to participate with its initiative. In particular, it provides free access to lecture recordings in audio or video form, along with the corresponding transcriptions, for a wide range of academic disciplines. Examples of OCW platforms are MIT OpenCourseWare and YALE OpenCourseWare; these are used in this thesis, as stated in Section 2.4 in Chapter 2. Clearly, an integral part of these projects is reliance on the use of audio transcriptions of the recorded lecture. However, the lecture transcriptions provided on these on-line platforms are delivered by commercial transcription services.

Manually transcribing this material is very tedious and costly, in terms of time and effort. For example, to manually transcribe a one-hour recorded lecture it requires at least 5 hours in the hands of qualified transcribers (Hazen, 2006). Alternatively, the work might take 10 hours by students enrolled in the course (Munteanu et al., 2008). Even when professorial transcribers are hired, the resulting transcriptions do deviate from the actual speech. Hazen (2006) observed that speech disfluencies such as filled pauses, false starts and partial words were removed or corrected in the transcriptions. Accordingly, there is a need to develop an automatic model to boost webcast lecture archives with high quality transcriptions, in order to improve access to these recorded archives and reduce human transcription time.

4.1.1 Motivation

Several approaches have been proposed to develop ASR systems for uncontrolled and diverse conditions such as lecture speech. One of the main challenges for ASR in such lecture conditions is the poor performance of the resulting lecture transcriptions. This is mainly

due to the mismatch between the language used in the lecture and the language models used in the ASR system. The great effects of language model mismatch come from the fact that lecture speech usually contains a high degree of spontaneity and topic-specific terminology (Furui, 2003, Glass et al., 2004). Thus, it is difficult to find other lecture sources to train the LM that match the style of the target lectures. Language models need to be constructed from other sources that are found in abundance. Hence, there is a need to accommodate the spontaneous speaking style of lectures, and to take into account the topic-specific terminology of the lecture subject. Given these conditions, the resulting LMs could be extremely useful for providing both accurate automatic transcriptions, and evidence for lecture segmentation and information retrieval. For example, Alharbi and Hain (2012) show that one can use the perplexities of adapted LMs to infer the general structure of the lecture from different disciplines.

Various methods have been proposed for LM adaptation. One of the most primitive approaches that shows effectiveness is the linear interpolation between two LMs. The key idea is to have small in-domain LMs (*e.g.,* lecture resources) and large out-of-domain LMs (*e.g.,* written text collected from the web), with an interpolation weight optimised on the development set. Such methods prove effective in decreasing WERs for many ASR systems. Another issue related to the evaluation of these sets of transcriptions is that it is not straightforward. The reference transcription in this case is approximate and not accurate (not verbatim). Most of previous work uses ground-truth transcriptions with time information to reliably evaluate the ASR outputs.

This chapter therefore tackles two important issues that have not been addressed before, with regards to the automatic recognition of OCW audio lectures. Firstly, a robust ASR system is built to generate automatic transcriptions. To speed-up the decoding process when relatively large interpolated LMs is used, a lattice rescoring using the full LMs is used and then extract the 1-best. Secondly, a new corrected version of the approximated transcriptions was produced to evaluate the ASR outputs. This was done by employing a lightly supervised alignment to alleviate the shortcomings of the existing imperfect transcriptions. This method attracts many researchers to align audio with imperfect transcription since it has two benefits: correcting the error in the transcripts and providing time information for each segment as well. This step is considered as pre-processing step to the scoring of ASR outputs. The resulting Word Error Rate (WERs) scores are approximate in this case due to the absence of the ground truth transcriptions. Having an accurate score of WERs is not important for our task (*e.g.,* metadiscourse tagging) as this was considered as a step in the pipeline as shown in Figure 2.4.

4.1.2 ASR-OCW System: Overview

This chapter presents a complete ASR system for academic lectures. The system is designed specifically to deal with OCW lecture resources. The AM is based on a DNN system on hybrid configuration, which has been shown to improve GMM-HMM models in many tasks (Hinton et al., 2012). The LM employed in this study is a set of three LMs, each of which was interpolated from in-domain and out-of-the-domain resources of over 600M words, with an interpolated weight optimised on a development set. All of the LMs possess a vocabulary of 36.2k unique words. Decoding was done using the Kaldi toolkit (Povey et al. (2011); see Section 4.3.1). Kaldi is a speech recognition toolkit consisting of a library, command-line programs and scripts (see Section 4.3.3 on how the decoding process is done). The LMs were pruned and converted in a Weighted Finite State Transducer (WFST; Mohri (2004), Mohri et al. (2002)) for decoding as required by the Kaldi toolkit. WFST is often used to provide a common and natural representation for context-dependency, pronunciation dictionaries, grammars, and alternative recognition outputs (Mohri et al., 2002). Optionally, lattices were rescored with the full LM to improve performance. Scoring references were corrected following a lightly supervised approach, such as the one defined in Task 2 of the Multi-Genre Broadcast (MGB) Challenge developed by Saz et al. (2015). The evaluation metric is the typical Word Error Rate (WER) (Klakow and Peters, 2002). The WER measure is computed based on the 1-best of the ASR hypothesis resulting from the decoding process (see Section 4.4.1), using the National Institute of Standards and Technology (NIST) scoring tool (NIST, 2009). The NIST toolkit is a well-known package in scoring speech recognition tasks. Again, the results obtained were approximate, as this task is just one step in the metadiscourse tagging approach. The actual evaluation of the ASR outputs will be presented in the next chapter when investing it with the metadiscourse tagging model.

4.1.3 ASR-OCW System: Contributions

The main contributions of the proposed work fall into the following aspects.

- Construction of automatic speech recognition for academic lectures for different disciplines.
- Creation and testing of several interpolated LMs using in-domain and out-of-the-domain resources.
- Evaluation of the ASR transcriptions by applying lightly supervised alignment to the approximated reference transcriptions.

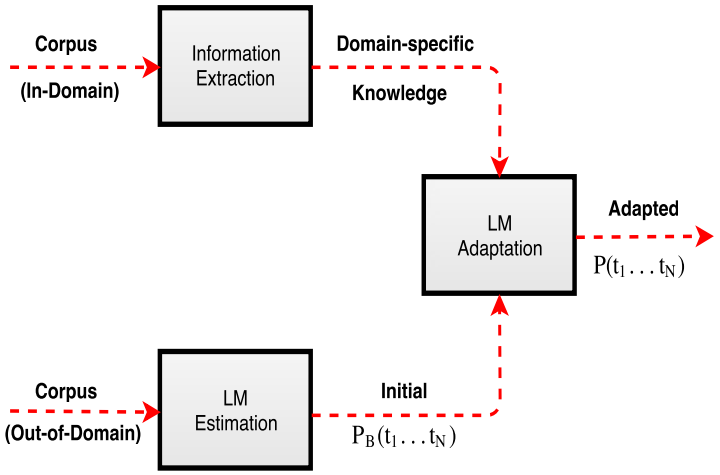


FIGURE 4.1: Architecture of language model (LM) adaptation. The figure is adapted from (Bellegarda, 2004).

4.2 Related Work

The focus on the development of the ASR-OCW system is mainly devoted to language model adaptation. Therefore, techniques for the statistical n-grams LM adaptation are reviewed in Section 4.2.1. Approaches to language model adaptation in the specific area of lecture processing are also discussed in Section 4.2.2.

4.2.1 Language Model Adaptation Techniques

Natural language displays different forms of variability, such as dynamic changes in the vocabulary usage in a given discourse, with respect to time (Bellegarda, 2004). In addition, the domain has an impact on the relevance of the word to the given discourse, which in turn yields different word statistics. For example, consider the relevance of the word sequence "interest rate" to an Economics lecture compared to a Physics lecture. Moreover, the syntactic forms of the word sequence cause another layer of variability in the language used. Finally, a speaker's style introduces a variability that may depend on the speaker's emotional status. Accordingly, the lexical, syntactic or semantic characteristic of the discourse in the training of an ASR model are different to those in the recognition stage. This causes a language model mismatch, which causes severe degradation in the ASR performance. LM model adaptation is often applied to reduce such mismatches.

More specifically, the task of LM adaptation involves two text corpora: an in-domain corpus (small) that is highly relevant to the target test set domain, and a background or out-of-the-domain corpus (large) that is gathered from different domains than the target domain.

A general architecture of the LM adaptation is described in Figure 4.1. The main idea is to dynamically update the initial estimate of the out-of-the-domain LM model according to the domain-specific knowledge.

Mathematically, given a sequence of N tokens t_q (where q, $1 \leq q \leq N$), the aim is to calculate an estimate of the LM probability:

$$P(t_1, ..., t_N) = \prod_{q=1}^{N} P(t_q|h_q), \tag{4.1}$$

where h_q represents the word's history at time q. By applying the Markovian assumption on the n-gram model, it becomes:

$$h_q = t_{q-n+1}, ..., t_{q-1} \tag{4.2}$$

Then, the conditional probabilities of $P(t_q|h_q)$ can be estimated by the well-known maximum likelihood estimate; the LM probability becomes:

$$P(t_1, ..., t_N) = \prod_{q=1}^{N} \frac{C(t_q, h_q)}{C(h_q)} \tag{4.3}$$

$C(h_q)$ and $C(h_q t_q)$ represent the counts of the token sequence h_q and $h_q t_q$ in the corpus (*e.g.,* A or B). In this context, it is important to note that any LM probabilities (*i.e.,* $P(t_1, ..., t_N)$) needs to be smoothed to prevent overfitting and to obtain a robust LM model.

The $P(t_1, ..., t_N)$ captures two different knowledge sources: in-domain corpus (A) and out-of-the-domain LM (B). The latter yields an initial estimate of $P_B(t_1, ..., t_N)$, which will be updated according to the relevant information from the former. There are various methods to perform such adaptation procedures. The most popular modelling techniques developed for academic lecture language models is the model interpolation techniques. These methods can be organised into four classes according to Bellegarda (2004): linear interpolation, back-off, cache model and mixture model. The cache model and mixture models can be thought as special cases of linear interpolation.

Linear interpolation is considered the most straightforward combination of two models, as required to compute the $P(t_1, ..., t_N)$, as follows:

$$P(t_1, ..., t_N) = (1 - \lambda)P_A(t_q|h_q) + \lambda P_B(t_q|h_q), \tag{4.4}$$

where $0 \leq \lambda \leq 1$ represents the interpolation coefficients. This parameter can be a function of the word history h_q, or can be estimated based on validation data. Alternatively, this can be estimated using the maximum likelihood criterion, using the expectation maximisation algorithm (EM; (Dempster et al., 1977, McLachlan and Krishnan, 2007)).

Back-off is performed from the out-of-the-domain LM to the in-domain LM, depending on the corresponding counts. This can be seen as a fill-up technique (Besling and Meier, 1995). An example of such implementation is below:

$$P(t_q|h_q) = \begin{cases} P_A(t_q|h_q) & if C_A(h_q t_q) \geq E; \\ \beta P_B(t_q|h_q) & otherwise, \end{cases} \tag{4.5}$$

where E is a threshold, and the back-off coefficient β is calculated to ensure that $P(t_q|h_q)$ is a true probability.

Cache Models are considered a special case of linear interpolation and are well suited for in-domain adaptation cases. This type of model was introduced for speech recognition by Kuhn and De Mori (1990). The intuition behind it is that words occurring recently in some text have a higher probability of re-occurring (Kuhn and De Mori, 1990). That is, the model utilises some words, say t_q in corpus A, to gain word statistics that cannot be computed from corpus B, then boosts their probabilities within the LM. In particular, this applies to the case of $n = 1$ of the previous interpolation model. For instance, the cache model for t_q is as follows (Clarkson and Robinson, 1997):

$$P(t_q|t_1, t_2, ..., t_{q-1}) = \lambda P_{cache}(t_q|t_1, t_2, ..., t_{q-1}) + (1 - \lambda) P_B(t_q|t_{q-2}, t_{q-1}), \tag{4.6}$$

In addition to the standard word tri-gram models, cache models can be interpolated with class-based, skip, and sentence mixture models (Goodman, 2001), as well as topic mixture models (Iyer and Ostendorf, 1999).

Mixture Models are another LM adaptation technique that captures the underlying topics of corpus A when estimating the LM (Bellegarda, 2004). Then, in some approaches it is used to boost the estimation of the background LM (from corpus B). One of the simplest approaches is based on linear interpolation. One assumes that tt_k is a set of topics that cover the underlying semantics of the background corpus B, and that the n-gram model of B consists of K sub-models, each of which has been trained on separate topics. This set of models includes an n-gram model trained on the whole corpus of B, in addition to small models trained on several portions of corpus B. Then, the mixture LM is defined by linearly interpolating these K n-grams to best match the adaptation data A, as follows:

$$P(tt_q|h_q) = \sum_{k=1}^{K} \lambda_{A,k} P_{B,k}(tt_q|h_q), \tag{4.7}$$

where $P_{B,k}$ represents the k_{th} pre-defined topic sub-model and $\lambda_{A,k}$ is the interpolation coefficient, which is estimated on corpus A.

In summary, most of the adaptation techniques presented above rely on the quality of the adaptation data used (*i.e.* corpus A). Thus, in the following section, various approaches are reviewed to gathering corpus A relevant to academic lectures, alongside a description of the techniques used within the framework of the ASR for academic lectures.

4.2.2 Adaptation Approaches in Academic Lectures

In general, most of the previous work in lecture processing in the area of LM adaptation relied particularly on constructing a fixed n-gram LM for lectures via linear interpolation of a domain-specific LM and an out-of-the-domain LM. The latter can be obtained from multiple sources, such as general conversational speech, meetings conversations, textbooks or web-based text. However, this is not the case for some studies, as will be discussed below.

Instead of relying on the dominant approach of linear interpolation of the in-domain and out-of-the-domain resources, Hsu and Glass (2006) attempted to develop an LM that best matched the topic and style of the target lecture by dynamically interpolating a generic style model with a topic-specific model related to the target lecture. The Hidden Markov Model with Latent Dirichlet Allocation (HMM-LDA) was used to acquire syntactic state and semantic topic labels for word instances in the training corpus. Then, an LM was constructed using these labels by extending the traditional bag-of-words topic models to n-gram statistics. The out-of-the-domain resources included only computer science (CS) related materials. It is not clear how general the model is when trained and tested on a variety of topics from different disciplines. To validate the effectiveness of the proposed approach of LM adaptation, a speaker-independent speech recogniser was used (Glass, 2003). Despite the appropriateness of their approach to a model that crosses style and topics of lectures, results show a relatively small reduction in WER (2.1%) when using the static topic model interpolation over the baseline of linear interpolation, which was built using the SRI Language Modelling Toolkit (SRILM; Stolcke (2002)). Further improvement was obtained by using adaptive interpolation of mixture components, which provided a further 0.3% reduction in WER.

In another line of research, Glass et al. (2007) performed LM adaptation by running a topic-adapted model. Instead of computing the n-grams statistics from HMM-LDA generated labels, n-grams statistics were computed from supplemental material. To be precise, the

adaptation procedure was done in two steps. First, a list of vocabulary was collected from the supplemental material (*e.g.,* journal articles, book chapters, and lecture slides) to be included in an existing topic-independent vocabulary. Then, the topic-independent n-gram statistics from an existing corpus of lecture material was integrated with the topic-dependent statistics of the supplemental resources, in order to develop a topic-adapted model. This approach was implemented as the previous approach, using the mixture language model capability of the SRILM Toolkit (Stolcke, 2002). As in the previous study, the approach was validated by using the speech recogniser developed for a telephony-based dialogue system (Glass, 2003). The baseline model was trained from a combination of transcribed lectures collected at MIT (1.3M words), in addition to data from the Switchboard corpus (3.1M words), and the MICASE corpus (1.7M words). The test set consisted of 6.1 hours of speech from 5 lectures about ASR, and the rest were public seminars given at MIT. The combination of acoustic and language model adaptation achieved a relative reduction in WER of 16% (from 33.6% WER to 28.4% WER).

In a follow up study to improve a real-time recognition system, Cerva et al. (2012) developed a client-server system for hearing-impaired students. The ASR module is considered an integral part of this system, providing a real-time recognition system for the highly inflected Czech language. To cope with the spontaneity of the lectures in Czech, unsupervised incremental speaker adaptation methods were used in developing the acoustic model. As in previous studies, there was an attempt to accommodate the challenges of building the LM, such as topic-specificity and spontaneity. The adaptation of the LM followed the same approach as the previous two studies of linear interpolation of in-domain and out-of-the-domain resources. The latter resources were extracted from several corpora, such as transcripts of spontaneous utterances, theses and web discussions. Then, linear interpolations were applied, where the weights were optimised on a development set using the SRILM toolkit (Stolcke, 2002). In addition to this domain-adapted LM, another general LM was constructed using all training corpora utilising a large lexicon. The proposed approach using these two LMs was evaluated on 18 hours of Czech lectures on Economics and Informatics. Results show that the domain-adapted LMs provide better results than the general one.

Bell et al. (2013) attempted to investigate the effectiveness of a neural network based LM, namely a factored recurrent neural network (fRNN), to improve the accuracy of the automatic transcriptions of lecture recordings. In addition to the standard LM described earlier, this was based on linear interpolation of in-domain and out-of-the-domain n-grams models. However, the difference is that the study applied a filter method based on cross-entropy (Moore and Lewis, 2010) to the out-of-the-domain in order to reduce the mismatch between the in-domain and out-of-the-domain corpora when developing the interpolated LMs. These resources are a small in-domain corpus from Technology, Entertainment, Design[1] (TED) and

[1] http://www.ted.com

a large set of out-of-the-domain news corpora. Then, the interpolated LM was implemented with a modified Kneser-Ney smoothed n-gram LMs (n $\in$ 3,4) using the SRILM toolkit (Stolcke, 2002). For the fRNNs LM, only a subset of the data was used, due to practical considerations, consisting of 30M tokens. The proposed approach yielded remarkable improvements for the TED Talk transcription task compared with the 2012 IWSLT evaluation campaign (Federico et al., 2012) from the University of Edinburgh, UK (UEDIN) (Hasler et al., 2012), and the National Institute of Information and Communications Technology, Japan (NICT) (Yamamoto et al., 2012). The results obtained using the fRNN LM are the best reported on for this task, achieving a relative WER reduction of more than 16% compared to the closest competing system.

Martínez-Villaronga et al. (2014) present a simple but powerful language model adaptation approach that is based on information retrieval. The approach was based on slide-based adaptation, and then compared with two strong LM baselines: n-grams based and slide-based interpolation LMs. The former was trained using a large collection of out-of-the-domain and in-domain documents, incorporating up to 46 billion words, while the latter used text slides extracted by optical character recognition (OCR), as described by (Martínez-Villaronga et al., 2013). These adaptation techniques were validated with an ASR system that consisted of two acoustic models: the standard Hidden-Markov-Model (HMM) and the Context-Dependent Deep-Neural-Network Hidden-Markov-Model (CD-DNN-HMM) approach (Seide et al., 2011). The word-list size was 200k words, using all of the out-of-the-domain corpora plus the in-domain vocabulary. The adaptation approach presented yield improvements on WER of up to 14% relative to the two competitive baseline LMs. In addition, the results show that part of this remarkable improvement was due to the use of the CD-DNN-HMM acoustic model. Despite the effectiveness of this approach, the are some limitations in this work as the resulting OCS may exhibit some errors that can be propagated in the LM used.

Akita et al. (2015) followed a similar approach to the previous study, in using the text of presentations slides for LM adaptation by extracting it using OCR. However, the proposed approach applied a filter method based on morphological and topical information to reduce the errors that result from the OCR, then performing a linear interpolation of the baseline LMs with the filtered text, alongside other resources chosen automatically from the text database. The baseline model is a tri-gram model originally trained on all lecture transcripts in the Corpus of Spontaneous Japanese (CSJ) (Furui et al., 2000). The size of the training texts was 7.7M words, and the vocabulary size was 37k words. The interpolation weights were optimised based on perplexity over the development set, which comprised of three lectures from the same collection as the test set. Further adaptation was obtained in the resulting LM by using the keywords in these filtered resources to improve word probability.

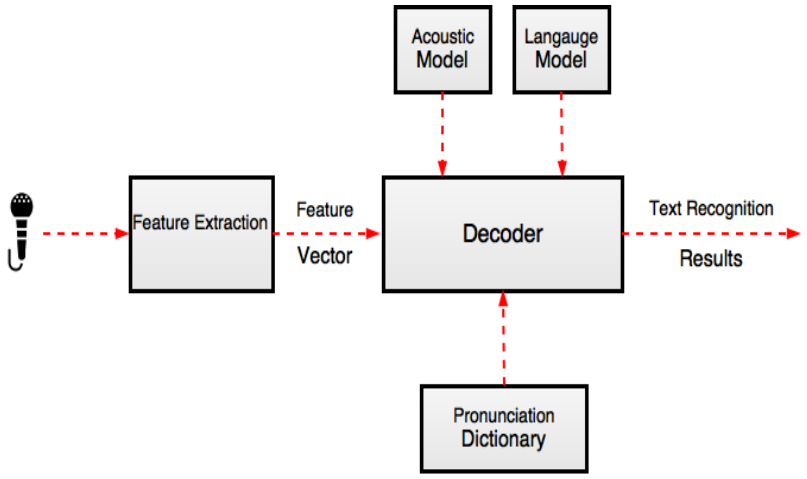


FIGURE 4.2: Standard architecture of statistical speech recognition.

Experiments on real lectures show significant improvements using this adaptation approach over the baseline.

In summary, the adaptation of a LM through linear interpolation of in-domain and out-of-the-domain resources is still the dominant approach in the specific area of lectures processing. Moreover, this approach is often treated as a strong baseline for more advanced LM, such as fRNNs. However, most of the previous studies share one similarity in developing the interpolated LM, which is finding relevant in-domain resources that best match the lectures in the test set. Some of these approaches rely on the use of text on slides or using OCR to extract the text if the slide is not available, which suffers from a lot of errors. Even if one applies some filter techniques on the extracted text in order to reduce the amount of errors, as seen above, this in turn greatly reduces the size of the resources available for LM adaptation. As an alternative approach, in this work an effective in-domain LM was constructed using transcriptions of real academic lectures from a wide-range of topics and disciplines. In this way, the resulting LM is a better match for the style and topics of the test set. This latter model was then interpolated with a huge background model that was developed using out-of-the-domain resources extracted from the web.

4.3 Lectures Transcriptions System

This section presents the work conducted to develop the ASR system. First a detailed description of how the acoustic models were developed is given. Then further descriptions of the language model used and the steps taken to conduct the decoding process are provided, as well as the description of the lightly supervised alignment model used to enable ASR scoring and evaluation. The purpose of this section is not to show how the best recognition performance can be achieved, but to explore the inherent properties of the lectures data,

Data	# Hours	# Talk/Lectures	# Speakers	# Segments
TED	289	1804	1804	164997
LLC	132	218	33	336392

TABLE 4.1: Data for acoustic model training

and to produce ASR transcriptions for the use in the next chapter. Thus, both acoustic and language modelling follow standard paradigms in speech recognition.

More formally, the aim of any ASR-OCW system is to find the most probable sequence of words T^* given the acoustic observations. The probability of the acoustic features $P(a)$ can be eliminated since it is the same for all, as follows:

$$\begin{aligned} T^* = arg\max_T P(T|a) &= arg\max_T \frac{P(a|T)\ .\ P(T)}{P(a)} \\ &= arg\max_T P(a|T)\ .\ P(T) \end{aligned} \tag{4.8}$$

The aim of acoustic modelling, therefore, is to estimate the parameters θ of the model to compute the probability of $P(a|T;\theta)$ accurately. The probability of the LM serves as guidance in the search process to interpret the acoustic input a. Figure 4.2 illustrates the process of decoding the most probable word hypotheses T^* for the given speech utterance. Initially the audio signal is sampled and processed to extract the acoustic features. Then these acoustic features are taken as input in the decoding process. More details about the strategies employed with respect to ASR system components are described in the following sections.

4.3.1 Acoustic Model

The acoustic model (AM) is arguably the main part of any speech recognition system. The AM estimates the probability of $P(a|T;\theta)$ of generating acoustic features for a given set of tokens T, and thus directly affects speech recognition quality, as seen in Equation 4.8. Note that the AM model is built using the Kaldi toolkit (Povey et al., 2011). Kaldi is a speech recognition toolkit consisting of a library, command-line programs and scripts. In addition, Kaldi uses Viterbi training for estimating AMs and several decoders for AM evaluation.

The main speech data used to build the model are demonstrated in Table 4.1. The TED talks data originated from 734 TED talks published before 31 Dec 2010. The duration of each talk is 15 minutes. The transcription is in the form of subtitles with rough segmentation, with segment durations of 3 to 5 seconds and time accuracy to the nearest second (W. M. Ng

et al., 2015). In addition, the Liberated Learning Consortium (LLC) is also used in training the AM. The LLC is an international research network dedicated to improving access to information through speech-recognition-based captioning and transcription systems. The corpus used is derived from a collection of 247 academic lectures on a wide range of topics, with a total of 150 hours of speech. Lecturers are mostly native English speakers from three main accent groups: North American, British, and Australian English. The length of individual lectures ranges from 38 minutes to more than one hour (Alharbi and Hain, 2012). The LLC reference transcripts includes timings per token. Hence, in order to get significant coherent segment, adjacent tokens without silence gaps were merged (Alharbi and Hain, 2012).

There are a number of studies that show hybrid neural network-HMM systems outperforming the state-of-the-art HMM-GMM systems for the task of large vocabulary continuous speech recognition (LVCSR), such as Dahl et al. (2012), Hinton et al. (2012), and Seide et al. (2011). These remarkable improvements can be explained by a number of points: the use of DNNs with many hidden layers, and with modelling context-dependent phone states, which results in a larger number of outputs classes. For this reason, the AM used is the Hybrid DNN-HMM model, which was developed by W. M. Ng et al. (2015). The Hybrid DNN-HMM model combines both Deep Neural Network (DNN) and Hidden Marokv Models (HMM). The inputs to the system were 5 contiguous spliced frames of Mel Frequency Cepstral Coefficients (MFCCs) features of 40 dimensions. The obtained features were produced using a linear discriminant analysis transformation of 117 spliced MFCCs features (from 13 dimensions with a context of 4 to the left and right and middle frame). Then, these features were transformed using a boosted Maximum Mutual Information (bMMI) discriminative transformation (Povey et al., 2008) to generate a more accurate target transcription for DNN training. The DNN model consisted of 6 hidden layers of 2,048 neurons, and an output layer of 3,830 triphone state targets. The target functions are State-level Minimum Bayes Risk (sMBR) (Gibson and Hain, 2006, Kingsbury et al., 2012), and Stochastic Gradient Descent (SGD) was used as the optimisation method.

4.3.2 Language Model

Although the AM is the main component in any ASR system, LMs also play a vital role in the recognition process. The LM assigns a probability estimate $P(T)$ for a sequence of words T. Usually, these probabilities are incorporated into the ASR search at an early stage and formulated according to a chain rule. The aim of $P(T)$ is to maximise the information about the next token t, given its history. For the purpose of improving the recognition results, two language models described in Alharbi and Hain (2012) and in W. M. Ng et al. (2015), respectively, are used in this study. After text preprocessing, several in-domain and

Data	**# Words**
In-Domain sources	
LLC	1.4M
BASE	1.5M
MICASE	600k
TED LECT	72k
Out-of-Domain sources	
Web(CHIL)	68M
Web(RT07)	40.5M
Web(Fisher-conv)	500M
Web(ami-rt05)	78M

TABLE 4.2: Number of words for different LM_1 resources in millions and thousands.

out-of-the-domain datasets were used not only to build these models via linear interpolation techniques, but also to create a lexicon suitable for lectures. Additionally, a further interpolation of these two LMs was exploited. Note that all of these LMs were implemented using Kneser-Ney discounting and standard back-off, and were trained with the SRILM toolkit (Stolcke, 2002). All these steps are described below in more detail.

Text Preprocessing: The preprocessing stage is designed to clean the text from formatting tags and normalise characters, expand abbreviations and normalise words, replace digits with their spelled out equivalent, remove punctuation and convert text into capital case.

Word-list Selection: The existing repository for in-domain materials of academic lectures in LM_1 exhibits around 36k unique words. This list was used in building the following language model.

LM_1: The first LM_1 was built using in-domain and out-of-the-domain resources, as described in Table 4.2. Aside from the LLC corpus, the in-domain consists of a collection of spoken lectures material that mostly matches the test set in style spontaneity and topic specificity: the Michigan Corpus of Academic Spoken English (MICASE)[2] consists of 1.8 million words of transcribed speech from a variety of events at the University of Michigan (Simpson and Swales, 2002); the British Academic Spoken English (BASE)[3] corpus contains transcriptions of 160 lectures and 39 seminars recorded at the Universities of Warwick and Reading. These are in addition to the TransLanguage English Database (TED LECT) (Lamel et al., 1994). The out-of-the-domain resources are the same as those used for meeting transcriptions in the AMI projects (Hain et al., 2005) and are partially based on web data collection (Bulyko et al., 2003).

[2] http://quod.lib.umich.edu/m/micase/
[3] http://www2.warwick.ac.uk/fac/soc/al/research/collect/base

Data	**# Words**
In-Domain sources	
TED Talks	3.17
Out-of-Domain sources	
News Commentary	0.9
Common Crawl	36.1
Gigaword	271.2
Europarl	10.8

TABLE 4.3: Number of words for different LM_2 resources in millions.

LM$_2$: The second LM_2 was also generated based on in-domain and out-of-the-domain resources, as shown in Table 4.3. The main corpus is TED Talks and the out-of-domain datasets are: News commentary v9, Common Crawl, Gigaword and Europarl v7. LM_2 is built on the full TED Talk data set and 25% or 50% of the out-of-the-domain data, making up to 322.2M words. The out-of-the-domain corpora were selected with the cross entropy difference (CED) criterion (Moore and Lewis, 2010). Sentences were ranked by their CED values and the 50% of the sentences with the lowest CED values were chosen from each out-of-the-domain corpus.

LM$_1$ + LM$_2$: is the combined version of the previously developed two LMs, namely LM_1 and LM_2. This was done by interpolating these two LMs using the same word list as before. Interpolation weights were optimised and tuned using maximum likelihood optimisation on the *dev* set. The *dev* set is about 29 lectures from different disciplines which formulate about 18 hours of speech. This set is taken from the 247 lectures collection of LLC and the rest is used to train the acoustic model as presented in Table 4.1 (and also further introduced in Section 4.3.1). Again, all of these LMs were implemented using Kneser-Ney discounting and standard back-off, and were trained with the SRILM toolkit (Stolcke, 2002). In addition, it is important to note that all of the three developed LMs (LM_1, LM_2, and $LM_1 + LM_2$) were considered in the recognition process, as explained in the next section.

4.3.3 Decoding Setup

As with the acoustic training process, the decoding was implemented using the Kaldi toolkit (Povey et al., 2011). In Kaldi, both LM and AM models are represented using a Finite State Transducer (FST). The decoding graph used is similar to the standard recipe described in (Mohri et al., 2002), where the decoding graph for the Weighted Finite State Transducer (WFST) is as follows:

$$HCLG = H \circ C \circ L \circ G, \tag{4.9}$$

where H, C, L and G represent the HMM structure, phonetic context-dependency, lexicon, and grammar, respectively, and ∘ is WFST composition. A WFST interprets the decoding problem as a beam search in a graph, where the task is to find the best path. When finding the best path, the input sequence corresponds to the state-level alignment, and the output is a word sequence constituting the utterance.

Since the underlying search space is large, to incorporate the LMs into the decoding pass, two decoding settings were considered and the results were compared;

- Lattices were generated for a pruned version of each language model: LM_1, LM_2 and the interpolated $LM_1 + LM_2$.
- Lattices were also generated using a highly pruned 3-gram for each of the three language models, and afterwards the lattices were rescored using the complete language models individually. For practical reasons, beam pruning was used to find the 1-best path instead of the full search.

It is worth mentioning that the decoding pronunciation dictionary was derived from a background dictionary containing pronunciation for over 136k words, based on the UNISYN dictionary and manual augmentation. The pronunciations of words not in the lexicon were automatically generated using the Phonetisaurus toolkit (Novak et al., 2012).

4.3.4 Alignment Model

This section presents the alignment model used to align the audio files with the reference transcriptions at the utterance-level. This step is important in the ASR-OCW system in order to get the time information that enables ASR scoring and evaluation. In addition, it is meant to correct imperfect transcriptions, as was the case with the current test set. Another advantage of having the time information for each segment is the ability to extract prosodic features in the reference transcription, as will be explained in Chapter 5. This is besides projecting the aligned reference to transfer the metadiscourse gold standard annotations to the ASR outputs, which is also described in Chapter 5 and Chapter 6. Similarly, the transferring of the reference thematic segmentation boundaries to ASR outputs is described in the application chapter, 7. These steps are important in order to evaluate the models on ASR outputs. In the following, a detailed description of the alignment system used is given.

The alignment system selected is based on a lightly supervised alignment that is more appropriate to align long segments than the standard Viterbi–based forced alignment, as studied by Hazen (2006) for lectures. In addition, it is useful in scenarios where the reference transcriptions are approximate and incomplete, such as is the case with OCW lectures. For

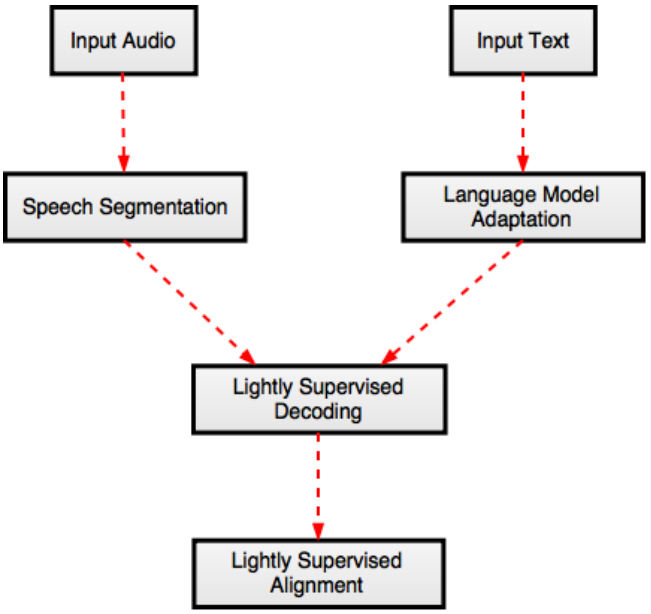


FIGURE 4.3: A standard approach to lightly supervised alignment. The figure is adapted from Olcoz et al. (2016).

that, the university of Sheffield system for Task 2 of the MGB Challenge was used; further description can be found in Saz et al. (2015). It is important to note that the decision to use this particular system was two-fold: first, there is a lack of lightly supervised alignment systems for lectures, as the task is more common in media archives than in lectures; second, this system is trained on media data of about 700 hours for the AM, with a total of 650 million words for the LM. This in turn explains the relatively good recognition performance obtained by this system, as it is comparable to the those trained on lectures in the ASR system, as will be shown in the results section (Section 4.4).

Most of the state-of-the-art lightly supervised alignment systems (Katsamanis et al., 2011, Lamel et al., 2002, Stan et al., 2016) share a common structure, as presented in Figure 4.3. The input audio is first segmented using Voice Activity Detection (VAD), which identifies the time boundaries for each speech segment. Text preprocessing steps are also applied to the reference transcriptions (*e.g.,* subtitles or lectures), similarly to those discussed in Section 4.3.2. Then, a background LM is adapted to this set of reference transcriptions. Afterwards, the speech segments obtained are processed in an ASR decoding step, which is often referred to as lightly supervised decoding. This latter step is accomplished using the adapted model and biased towards texts that are matched with the approximated reference transcription, ignoring the unmatched ones. This decoding step can be far more complex, such as the use of multiple decoding passes and speaker adaptation. Next, the generated transcripts hypothesis provided by the lightly supervised decoding stage is aligned with the approximated reference transcription. This step is often implemented through recursive dynamic programming approaches in which word sequences from the reference transcriptions

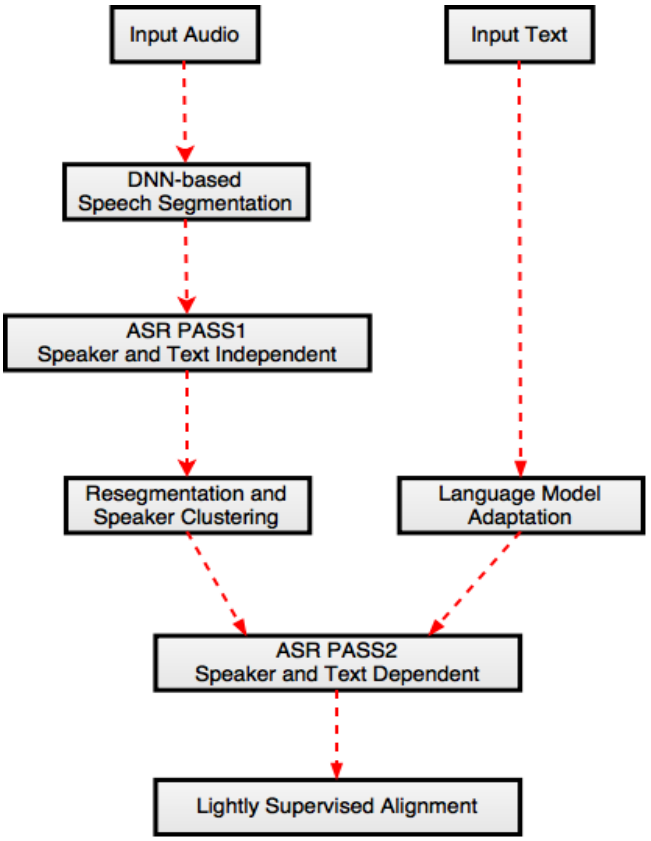


FIGURE 4.4: The MGB lightly supervised alignment system framework. The figure is adapted from Olcoz et al. (2016).

are assigned to the speech segment. This is done based on how well they match the output of the previous decoding stage by using, for example, Levenshtein distance (Katsamanis et al., 2011). The output of this stage is a set of speech segments that contain words from the original reference transcriptions. Finally, to determine precise word-boundaries for this output, a second alignment can be applied.

The structure of the lightly supervised alignment system used has a similar structure to the typical lightly supervised systems described above. Moreover, it follows the stages taken by Saz et al. (2015),, as described in Figure 4.4. The first stage of the system is the lightly supervised decoding, followed by lightly supervised alignment as the second stage. The lightly supervised decoding stage consists of several steps. First, a DNN–based speech segmentation is used to identified segments of the lecture speech. Then, an initial transcription is produced for these segments using an independent DNN–HMM system (Hinton et al., 2012). This system is trained on 700 hours of speech data and implemented using Kaldi tools (Povey et al., 2011). The decoding in this stage was done using a background language model trained with a total of 650 million words, using TV subtitles from 1979 to March 2008, again using the SRILM toolkit (Stolcke, 2002). The output of this stage is used in the second decoding stage after performing resegmentation, speaker clustering, and speaker adaptation.

Utter 1	Ground Truth	Demand is how much something wants someone wants something
	Approximated Transcripts	Demand is how much someone wants something
	Lightly Supervised Alignment	Demand is how much something wants someone wants something
Utter 2	Ground Truth	On the one hand I might you know that could have some pros and cons
	Approximated Transcripts	On the one hand, that could have some pros and cons
	Lightly Supervised Alignment	On the one hand I might you know that can have some pros and cons

TABLE 4.4: Examples of two utterances processed by the alignment system from one Economics lecture. For each utterance, the table shows the ground truth transcription, the approximate transcript provided by OCW, and the output of the alignment process.

The second decoding is based on a DNN–GMM–HMM system trained again on 700 hours of speech, using TNet (Veselỳ et al., 2010) and HTK (Young et al., 2006). However, the language model in this stage is interpolated using both the background model and lectures LM, with interpolation weights equally split at 0.5 for each component ($\alpha = 0.5$). The final stage is to align the lecture transcriptions in a recursive lightly supervised alignment process.

Table 4.4 shows an example of the output from this alignment system for two utterances. In both utterances, the human annotator for the approximated transcript removed speech errors produced by the lecturer. In utterance 1, the output of the alignment system matches the ground truth, whilst in utterance 2 there are some differences, such as the use of "can" instead of "could". This may be the result of using different language model contexts.

4.4 Experiments and Results

This section aims to presents the experimental setup selected to evaluate the outputs of the ASR-OCW recognition system. The results are presented along with discussion and conclusions.

4.4.1 Experimental Setup

Datasets

This section explains the datasets used for evaluating the ASR-OCW system, which are the main academic lecture course datasets used throughout this thesis, as fully described in Table 2.4 in Section 2.4. The same set was also used in conducting the annotation experiments, as shown in Chapter 3. This collection is composed of around 106 hours of academic speech from two different disciplines: Physics and Economics. However, this is different from the datasets used for training the ASR-OCW system and the lightly supervised alignment system, as described in Sections 4.3 and 4.3.4, respectively.

Evaluation Metric

To evaluate the output of the ASR -OCW recognition system, Word Error Rate (WER) standard metric is used. The WER measure is computed based on the 1-best of the ASR hypothesis resulting from the decoding process. That is, by comparing the ASR hypothesis with the reference, which is the output of the lightly supervised alignment system. This yields a score in the form of the percentage of errors. The types of errors that can occur when aligning the hypothesis with the reference are categorised as follows:

- Substitution: a word that is misrecognised.
- Deletion: a word in the reference but not in the hypothesis.
- Insertion: the recogniser inserts a word that is not in the reference.

After identifying these errors, the equation used to compute the WER is as follows:

$$WER[\%] = \frac{\#substitutions + \#deletions + \#insertions}{\#words(reference)} * 100\% \tag{4.10}$$

This equation of the WER is an error function, hence, the lower the value the better. For example, a value of zero indicates that the hypothesis transcriptions match the reference ones, whereas, a value of 100 indicates that there is a complete mismatch between them. It is worth noting that this Equation (4.10), indicates that the WER values could exceed 100%, if for instance the hypothesis transcriptions contain many insertions.

Systems implementation

The two systems that were used in this chapter, namely the ASR-OCW system and the lightly supervised alignment system, were implemented based on the Resource Optimisation Toolkit (ROTK). The ROTK was developed by the group at the University of Sheffield (Hain et al., 2012). It provides an asynchronous execution of the system modules using a grid computation infrastructure. In particular, the ROTK system is defined as a set of linked modules that transfer data of specific types in slightly similar ways, as described in Figure 4.4. In addition, the metadata is used to organise data flows as in the diagram. It also allows parallel execution of the module by splitting the module into subtasks and keeping track of the dependencies between these subtasks. Each module submits jobs on a grid system using the Sun Grid Engine (SGE). The ROTK system allows for simple repeatability of the experiments, as the same graph can be executed on different datasets, such as development and evaluation sets.

	Physics				**Economics**				
LM	Sub	Del	Ins	WER	Sub	Del	Ins	WER	**Overall**
LM_1	12.15	6.75	9.8	28.70	10.60	10.10	7.60	28.25	28.43
LM_2	13.25	7.00	9.80	30.00	10.95	10.45	7.20	28.55	29.28
$LM_1 + LM_2$	12.15	6.80	9.75	28.65	10.55	10.20	7.45	28.10	28.38

TABLE 4.5: Word error rate (WER in %) for the two disciplines for the LM used.

	Physics	**Economics**	
	PPL	PPL	Average
LM_1	152.88	154.00	153.44
LM_2	272.36	304.81	288.59
$LM_1 + LM_2$	158.37	155.71	157.04

TABLE 4.6: Perplexities for the test set from Physics and Economics courses, using the three language models: LM_1 and LM_2 and the interpolation of the two, $LM_1 + LM_2$.

4.4.2 Results

The aim of this section is to analysis the ASR performance in the ASR-OCW system using the three LMs: LM_1, LM_2 and $LM_1 + LM_2$. Further, two different decoding settings for LM_1 were investigated: Pruned LM, and Pruned and rescored. Meanwhile, only one decoding setting was considered for both LM_2 and $LM_1 + LM_2$ due to the low performance in terms of WER of these LMs compare to LM_1. Furthermore, a distinction is made in representing the results between the two different disciplines, Physics and Economics, in order to analyse the results accordingly.

Table 4.5 shows the results of ASR models using the three language models in both disciplines. Note that in order to fit each LM into the WFST, these three LMs were fully pruned for the recognition task. Unsurprisingly, the best results were achieved by using the LM_1 with a WER of 28.60% and 28.25% in Physics and Economics, respectively. This can be attributed to the fact that the LM_1 is trained using real lecture resources from different disciplines that share some characteristics with the word n-grams in the test set. However, the results in Economics when the LM_2 (28.55%) are not that different from those obtained by LM_1 (28.25%). This is in contrast to the results for Physics lectures using the same model. This can be explained by the fact that Physics lectures contain some specific terminology that does not occur in LM_2. Although the results obtained from $LM_1 + LM_2$ for Economics are better than those achieved by LM_2, they do now show any improvement compared to LM_1 on Physics lectures.

To have more insight into the quality of the LM on the test set, perplexity (PPL) is used. PPL is a measurement of how well a probability distribution or probability model predicts a sample (Bahl et al., 1983, Jelinek et al., 1977). A low perplexity suggests that the probability

	Physics				**Economics**				
LM	Sub	Del	Ins	WER	Sub	Del	Ins	WER	**Overall**
LM_1	11.85	6.75	9.70	28.30	10.40	10.10	7.50	28.05	28.18

TABLE 4.7: ASR results using pruned and rescored language model (LM_1).

	Physics				**Economics**				
LM	Sub	Del	Ins	WER	Sub	Del	Ins	WER	**Overall**
LM_1	15.90	9.00	9.30	34.20	15.90	13.55	6.65	36.10	35.15

TABLE 4.8: ASR results using BBC acoustic and language models that were used originally for in the alignment system.

distribution is good at predicting the sample. The test sets are the Physics and Economics lectures from MIT and Yale OCWs. The results of calculating perplexity on the test set selected for these three models are presented in Table 4.6. Not surprisingly, the lectures-based language model LM_1 performs far better than the other LMs, with relatively similar PPL in both disciplines: 152.88 and 154.00 in Economics and Physics, respectively. These results are complementary to those obtained in the recognition using WER, as explained above. On this basis, the results obtained by LM_1 are further improved by performing lattice re-scoring using the full LM on the pruned LM. Table 4.7 show that the results in both disciplines improved when using this setup over the LM. The WER results obtained in this setup were 28.18%, compared to the best results reported before (28.43%).

For the sake of comparison, the media models (both AM and LM) that are part of the lightly supervised system were used here as well, but for recognition purposes. Table 4.8 shows the results of this setup. Despite the fact that there is a data mismatch between the data that the model trained on (media) and the evaluation dataset (lectures), the media models are able to recognise lectures from both disciplines. However, the results obtained are worse than the above results. For example, for Physics lectures the model achieved 34.20% WER compare to 28.30% obtained from the lectures-based model. Similarly, for Economics lectures the media-based model scored 36.10% compare to the best results obtained from the lecture-based model (28.05%).

4.4.3 Discussions

The best recognition results were obtained when both the AM and LM trained on data with similar behaviour as the test set. In other words, when the models trained on in-domain datasets as the test set they could be expected to provide better results. This observation is very common in any ASR recognition task. However, this not always the case, as it has been shown that the models (AM and LM) trained on out-of-the-domain data (media collection)

are able to decode the lecture dataset. Thus, by augmenting both in-domain and out-of-the-domain data, not only for the language model as described, but also for the acoustic model, further improvements in results may be achieved.

Unfortunately, the results obtained in this chapter in terms of either PPL or WER cannot be compared with previous studies, as different datasets were used. Nevertheless, analysis of previous work based on general similarities with respect to the datasets used and LM adaptation techniques, may verify the approach taken in this chapter. That is, among the previous studies, the baseline LM model proposed by Hsu and Glass (2006) exhibits an equivalent approach in LM adaptation of the interpolation of in-domain and out-of-the-domain resources. The reported PPL of the current test set is slightly better (153.44) than the one reported in their study (154.4). In particular, their test set is very small (10 lectures) and covers only computer science topics, whilst the test set used in the ASR-OCW system contains 106 lectures from two different disciplines, and covers a wide range of topics. In another similar study, proposed by (Glass et al., 2007) a similar approach was used in developing the interpolated LM, and they used a similarly sized vocabulary list of 37.4k words, compare to the 36.2k word in this study. However, the study did not report the PPL; instead the out-of-vocabulary (OOV) rate was used to quantify the improvement in the WER. The OOV on their test set after adaptation was 0.64; the average OOV rate on this study is 1.8 in both disciplines. This can be attributed again to the fact that the size of the test set used in (Glass et al., 2007) is small – only 5 lectures compared to 106.

Furthermore, it seems adding either more in-domain or out-of-the-domain data for training the LM can improve the WER results. For example, LM_1 has more words, around 690M, compared to only 322M in LM_2. However, combining multiple n-grams may generate a very large number of parameters for the model, which is costly to handle in the decoding process (Bulyko et al., 2003). In addition, these long-span LMs cannot be represented using WFST. Thus, two approaches were considered: first a pruned LM is used instead of the full LM. The impact of the latter approach on LM_1 was very important in the ASR-OCW recognition system compared to the pruning approach. Such a modelling strategy proved effective in a number of speech recognition tasks dealing with long-span LMs, such the study proposed by Kombrink et al. (2012).

There are areas for improvements in the current ASR-OCW system, in particular with regards to modelling long span LMs such as Neural Networks (NNs), which can be expected to provide further improvement. For example, Sundermeyer et al. (2014) combined long-span neural network language models with lattice rescoring and decoding. However, the results obtained by the ASR-OCW recognition system are already quite good for the task of MDT tagging, as will be described in Chapter 5. In particular, this study investigates how the MDT model would treat errorful transcriptions that resulted from ASR. Improving

the recognition model in the ASR-OCW by focusing on the LMs would be useful for future research. Another important direction for future research would be the development of lightly supervised alignment system training on in-domain data (*i.e.,* academic lectures), and integrating that alongside the ASR components in the ASR-OCW system.

4.5 Conclusion

This chapter has presented the second stage of the metadiscourse tagging approach developed for academic lectures. In the previous chapter, a corpus of metadiscourse was developed for academic lectures using reference transcriptions. The aim was to investigate the occurrences of metadiscourse phenomena in academic lectures in two disciplines: Physics and Economics. The corpus will be used in the development of the metadiscourse tagging model in the next chapter, 5. However, it is also interesting to investigate the performance of the tagging model on automatic transcriptions, which is a challenging task as the transcriptions contain considerable errors.

In this chapter, therefore, an automatic speech recognition system was presented for challenging types of data, such as real academic lectures. The system consists of several main components, including an acoustic model (AM), language model (LM), decoding process and an alignment process. The focus of this chapter was on the development of the LM by linearly interpolating several in-domain and out-of-the-domain resources. The in-domain set included the transcriptions of real academic lectures from different disciplines, whilst the out-of-the-domain set consists of a large set of web-based resources. The interpolated weight was optimised on a development set of lectures. To evaluate the resulting ASR output, a lightly supervised alignment system was applied on the reference transcriptions, in order to produce time information for each speech segment. The experiment results show that the proposed LM outperforms the baseline LM, and that this is consistent across disciplines.

Chapter 5

Exploring Features for Metadiscourse Tagging with SVMs

The previous two chapters presented the pre-request that is needed to conduct the task of metadiscourse tagging. A corpus of metadiscourse within academic lectures was built up at two levels of granularity (specific and generic tags) and across two different disciplines, using manual transcriptions. In addition, an ASR system was developed to produce automatic transcriptions for the same set of academic lectures as used before. This chapter describes the baseline approach to automatically classifying utterances with metadiscourse tags. The proposed metadiscourse tagging model is based on a multiclass classifier based on SVMs (MDT-SVMs), and a set of sparse textual features combined with dense features such as prosodic cues. Experiments with both specific and generic sets of metadiscourse tags indicate the effectiveness of this approach. The use of feature combinations in particular words, lemma and POS tags n-grams, and prosodic cues gives the best classification results. Results also show that domain knowledge has an effect on the classification results. Further, when classifying some finer-level tags (specific), the model performed poorly. This can be attributed to the sparsity problem that is often encountered with similar related tasks. To investigate the model performance on ASR outputs, a strategy was followed of transferring the gold standard tags from the reference transcriptions to the automatic transcriptions. However, as expected, testing with ASR outputs showed a decrease in model performance across both disciplines.

This chapter is structured as follows: Section 5.1 introduces the topic of this chapter, combined with the motivations and contributions of the presented work. Section 5.2 describes the previously introduced feature types that have been used in related discourse coding tasks. Two sets of these effective feature types are also used here, to classify metadiscourse tags in academic lectures, as described in Section 5.3. The performance of this model is measured

using commonly known metrics – precision, recall and F-measure (Rijsbergen, 1979) – and an extensive analytical study is conducted using several test cases, including the use of ASR outputs, in Section 5.4. A concluding discussion is set out in Section 5.5.

5.1 Introduction

The typical modelling approach in most discourse coding related tasks, including metadiscourse tagging, is to conduct the task as a text classification task, which involves two components: a set of hand-engineered sparse (or high dimensional) features, and a classification model such as SVM. Despite the simplicity of this approach, it achieves success in most of discourse coding classification tasks and is often treated as the baseline model for comparisons with more advanced models. For example, for the task of dialogue acts tagging, (Fernandez and Picard, 2002), (Stolcke et al., 2000), (Surendran and Levow, 2006), and (Webb et al., 2005) used several types of features, lexical, syntactic and discourse cues, along with low-dimension features such as prosodic cues, with different classification models. For the task of metadiscourse tagging, Correia et al. (2014a) utilise such an approach to the task using only lexical, syntactic and positional features, which showed its ability to identify metadiscourse tags in Ted Talks with a decision tree as the classification model. In addition, both lexical and syntactic features have been used for the task of speech acts labelling in email (Cohen et al., 2004, Qadir and Riloff, 2011).

The main commonly used features in related text classification tasks are lexical features, in particular the frequency of n-grams in the corpus. Other features can be used as well, by annotating the text with particular types of features, such as the use of POS tags or named entity recognition (NER) tags. The typical approach to representing these types of features involves two steps: construct a dictionary of words by exploring the corpus and then construct a term-document (or term-utterance in the current case) frequency matrix. This approach is based on a bag-of-words model (Salton and McGill, 1986), where a text (*e.g.* an utterance) is represented as the bag (multiset) of its words, disregarding grammar and word order. The hypothesis here is that the vector in a term-document matrix captures to some degree an aspect of the meaning of the target document (Turney and Pantel, 2010). That is, the occurrences of words in a document tend to show the relevance of the document to a query or a target tag (Salton et al., 1975). Despite the popularity of this approach in related classification tasks, sometimes the high dimensionality can lead to overfitting problem. Several techniques have been proposed to solve this problem, in order to improve the generalisation and avoid the overfitting problem. For example, feature selection methods can be applied to reduce the dimensionality of the feature space.

However, feature selection assumes that there are some irrelevant features, and tries to determine these set of features to remove them. Feature selection may result in loss of information (Joachims, 1998). Another way to avoid high dimensional input spaces is through the use of classification models (such as SVMs) that are able to work with high-dimensional data. That is, SVMs can learn independently of the dimensionality of the feature space, by measuring the complexity of hypotheses based on the margin with which they separate the data, not related to the number of features. This means that one can generalise even in the presence of very sparse features. Another important property of SVMs is that they allow easy integration of sparse high-dimensional text features and dense low-dimensional features, such as acoustic features.

5.1.1 Motivation

The use of acoustic-based features such as prosodic cues has well-demonstrated effectiveness in various spoken-language understanding tasks, especially when combined with sparse textual-based features. Prosodic cues are known to be relevant to discourse structure across languages, and can therefore play a vital role in various information extraction tasks (Shriberg et al., 2000). The inclusion of prosodic cues along with other high-dimensional textual features, in particular pitch and pause duration, has shown great success in discourse related tasks, such as dialogue acts, in a number of different domains (*e.g.* meetings, conversational speech) (Shriberg et al., 1998, Stolcke et al., 2000). A previous study on metadiscourse tagging of TED Talks, proposed by Correia et al. (2014a) has shown that n-gram based features, along with other textual features such as POS tags, were suitable for the metadiscourse classification task, but that study did not explore the usefulness of prosodic cues. Therefore, for the task of metadiscourse tagging for academic lectures, it would be interesting to explore whether the inclusion of prosodic features is complementary to textual features, or may improve the model performance, or have no effects at all.

The overall aim of this work is to operate on ASR output, and testing the model on the automatic transcriptions is expected to decrease the model's performance. However, it would be interesting to know how much degradation one can observe using such an approach. Additionally, the work in this chapter is aimed at investigating the effectiveness of SVMs for the metadiscourse tagging task using textual-based and acoustic-based features. Previous work has not studied the effect of domain knowledge (*i.e.* discipline) on the tagging model, as it was applied in presentation-style domains such as Ted Talks, where the discussion is more about general topics, compared to more specific topics that used domain-specific terminology, such as university lectures. The annotated corpus in this thesis exhibits two disciplines, namely Physics and Economics; in this chapter a distinction is therefore made in modelling these two disciplines, in order to observe the effects of domain-knowledge.

5.1.2 MDT-SVM: Overview

This chapter presents the first modelling approach to the metadiscourse tagging task, using the annotated corpus of academic lectures as previously described. As with most utterance-level classification tasks, the approach involves two stages, the first related to the feature set used, and the other related to modelling procedures. For the feature set, a number of different feature types and a combination of these features were used and organised under two categories: textual features and prosodic features. Among textual features the word, lemma, POS tags n-grams and positional features were all examined for the task. For the prosodic features, the most prominent cues, such as F0 and pause durations (PD), were extracted using the speech-to-text alignment system described in detail in Chapter 4. To investigate the model performance on ASR outputs, the ASR model developed in Chapter 4 was used, with a performance of around 28% WER in both disciplines. A strategy was followed of transferring the gold standard tags from the reference transcriptions to the automatic transcriptions. To perform the metadiscourse tagging task, an SVM was used and for the purposes of this study the model is referred as MDT-SVM. However, an SVM is a binary classifier and the task of metadiscourse tagging requires a multiclass classification model. Thus, an extension of this basic model was used, where the task was decomposed into multiple classifiers (*i.e.* one classifier per class), according to a one-versus-all (OVA) approach. The commonly used metrics – precision, recall and F-measures – were chosen for evaluation.

5.1.3 MDT-SVM: Contributions

The main contributions of the work fall under the following categories:

- Developing an automatic approach for modelling the metadiscourse tagging task within academic lectures, using manual engineering features and SVMs.
- Investigating different lexical and prosodic features, or a combination of these, for the task, using this approach.
- Investigating the effects of using imperfect transcriptions resulting from ASR.
- Investigating the effects of domain-knowledge on the classification task.
- Analysing the classification performance at two levels of tag granularity: generic and specific.

5.2 Related Work

Since part of the scope of this thesis is discourse-related tasks, existing classification methods for modelling these tasks based on a combination of features are also reviewed. These features can be categorised into two classes: textual-based tasks and acoustic-based tasks. Section 5.2.1 discusses the recent literature about the use of textual-based features in related discourse tasks (but does not pretend to give a complete coverage to all works, especially for tasks other than metadiscourse). Section 5.2.2 introduces a short survey of several approaches that have used acoustic-based features.

5.2.1 Textual-based Features

The nature of metadiscourse tagging requires capturing different expressions that indicate different discourse functions. For this reason, prior research on metadiscourse classifications for both written and spoken discourse has been based largely on human transcriptions, and so has focused on textual information. However, there are three approaches in modelling metadiscourse: supervised, unsupervised and a combination of the two, as discussed in Section 2.2.2. This section provides a short survey of textual-based features used in these different approaches, while the main focus is on features for supervised models. The studies included here cover both spoken and written discourse.

One of the prominent textual features used in most approaches for modelling metadiscourse are word n-grams. For example, early work on metadiscourse looked for the most frequent word n-grams, where $1 <= n <= 6$, using the features to mark certain phrases that indicate rhetorical functions in scientific articles from two different disciplines: computational linguistics and medicine (Teufel, 1998). These n-grams are then filtered using a seed lexicon to keep the expressions that contain at least one of the words of the seed-lexicon or variants of it. Finally, a set of n-grams was compiled and counted from the entire corpus, to compare the filtered phrases with the annotated ones. In a follow-up study by Abdalla and Teufel (2006), an unsupervised method is proposed to find similar variants, by bootstrapping seeds from within the phrases. The method hypothesised that each metadiscourse expression contains at least two concepts whose syntax and semantics mutually constrain each other. The method shows effectiveness in detecting certain phrases that have similar patterns to the seed phrases.

Other approaches rely on handcrafted rules to capture the expressions patterns. For example, Madnani et al. (2012) attempt to find the phrase patterns by asking experts to compile a list of 25 hand-written regular expressions that match occurrences of shell text

in argumentative student essays. These patterns were produced by compiling a list of n-grams (where $1 <= n <= 9$) extracted from the annotated essays. Wilson (2012), however, used previous experience to find a set of words that act as indicators of metalanguage, such as meaning, sentence, or symbol. The main purpose of this metadiscourse extraction task studied by Wilson (2012), was to use these sets of words, generated from a rule-based strategy, to find candidate sentences prior to annotation. The methods in these two studies share some similarities with a method proposed by Riloff (1993) for learning information extraction (IE) patterns. The method uses a syntactic parse and a set of 1500 filled templates to learn context in terms of lexico-semantic patterns, and a lexicon of semantic features for roughly 3000 nouns. The main drawbacks in this set of studies is the labour required to develop these set of rules, which is both time consuming and costly.

Conversely, Wilson (2013) develops a two-stage process, namely detection and delineation, to automate the process of identifying metalanguage within the written text of the created corpus, as discussed before. The first stage aims to find sentences that contains instances of metalanguage. The authors used a mixture of stemmed and unstemmed words, *uni*-gram and *bi*-grams. Then, an improvement over the baseline approach was conducted by ranking those n-grams over the training set using information gain. They subsequently used only the top ten ranked list in the classification model. The results show that the performance of this approach roughly matches the performance of the inter-annotator agreement. In the second stage, an approach was developed to delineate sequences of words mentioned by a metalinguistic statement. The study manual examined the corpus to find sequence patterns between meta-words and mentioned language, such as noun apposition using TRegex search strings (Levy and Andrew, 2006) and semantic roles were explored similarly using the Illinois Semantic Role Labeler (SRL; Punyakanok et al. (2008)) for the automatic delineation of mentioned language. Although the overall results of the approach are preliminary, its accuracy shows promise for future development.

The previously discussed approaches required manual interventions in order to find the metadiscourse patterns. These are important when the task requires finding the exact sequence of words that compose the metadiscourse expressions. However, when the task is to find whether the sentence contains an instance of the metadiscourse, one can use a set of n-grams as a feature set, as studied by Wilson (2013), for the detection task. Correia et al. (2014a) show that the set of n-gram features is not only able to detect the phenomena in the sentence but is also able to differentiate between four different categories for academic speech. The work in this study is quite similar to the current study; however, it also shows that these n-grams features are able to differentiate between the expressions of a set of 19 different categories. Similar lexical approaches are also found in different research areas, such as word sense disambiguation (Pedersen, 2001), sentiment analysis (Abbasi et al., 2008, Pang et al., 2002), or feedback localisation (Xiong and Litman, 2010). The main intuition in all

of these studies is that words can be used as indicators of the presence of the phenomenon being studied. For example, certain words in sentiment analysis are associated with positive opinions, while others are neutral and others have negative tones.

To further improve the classification task based on word n-grams, Correia et al. (2014a) used POS n-grams to capture the metadiscourse phrase patterns for different tags. In particular, they used the presence of POS n-grams in the sentence. The POS representations convey grammatical information and thus are often used in other style-based categorisation studies. For instance, Lioma and Ounis (2006) examined the distribution of POS blocks in language by analysing their frequency in documents, under the intuition that the more frequently a POS sequence occurs, the more salient information its contains. More specifically, the hypothesis was that the distribution of POS n-grams in a corpus can indicate the amount of information they contain. To prove the effectiveness of this approach, the study tested this hypothesis in the context of information retrieval, and confirmed that high frequency POS n-grams are typically content rich, and that removing content poor POS n-grams from search engine queries results in an improved model performance. The common idea between all of these studies is that the syntactic forms of word n-grams can be used as indicators of the presence of the phenomenon being studied.

In addition, other types of features can be combined with traditional features (*e.g.* word n-grams) in order to improve the model performance. For example, Correia et al. (2014a) investigated the effectiveness of adding position and length information to the classification model. Examples of these features include length of the sentence, position of the sentence in the talk, and distance from the last occurrences of the metadiscourse category, in terms of the number of sentences. Results show that these did not add any improvements for the classification task across categories. However, in related tasks such as relation classification and sentiment classification, such features provide great improvements in the model performance. For instance, in a relation classification task, which is defined as assigning relation labels to pairs of words, position features are useful in encoding the relative distances to the target pairs (Zeng et al., 2014b). Moreover, Pang et al. (2002) show that the position of the word is useful in predicting the sentiment of movie reviews. The study hypothesis was that there is a common structure for a movie review, in that it begins with an overall sentiment statement, proceeds to a plot discussion, and concludes by summarising the author's views. Hence, the study attempted to approximate the positions of the words with respect to this structure. A similar study conducted by Kim et al. (2006) found that the average sentence length is related to the structure of user reviews and is useful in predicating the helpfulness of such reviews.

5.2.2 Acoustic-based Features

Previous work on metadiscourse tagging did not explore acoustic-based features; however, this study deals with a spoken corpus. The use of prosodic features has been well studied in the context of other related discourse tasks, particularly for discourse segmentation and dialogue act tagging. Prosody means here information about the temporal, pitch, and energy characteristics of utterances, independent of the words themselves. Hence, prosody can provide complementary information to the word sequence, and is thus a valuable source of additional information for the task at hand. Moreover, with the high WERs that result from ASR, prosody may give more robust features than textual information, as its facets are relatively unaffected by word identity (Liu et al., 2006). The following section explores the pioneering works in the literature that exploit prosodic features for dialogue act tagging.

Studies into the use of prosodic features for dialogue act tagging can be organised according to the type of prosodic representation investigated (categorical or continuous). For example, Black and Campbell (1995) used categorical representation of prosodic features for a dialogue act tagging task in text-to-speech synthesis, for use in a speech translation system. In particular, the study proposed a multi-level intonation system, which produces a fundamental frequency F0 contour, based on input labelled with high-level discourse information, including speech/dialogue act type. This F0 contour was labelled to describe the contour in terms of rise, fall and connection. The main aim here was to improve the speech synthesis system using the parameters of the intonation system. Results show that this approach is able to distinguish speech/dialogue act classes well. In addition, the authors concluded that some prosodic features work well for certain speech classes. Similar findings have also been presented in this regard by Kompe et al. (1997), who also found that prosodic models may help to capture the following typical characteristics of some dialogue acts:

- a falling intonation for most dialogue act of type statements.
- a rising F0 contour for some questions (particularly for declaratives and yes/no questions).
- a continuation-rising F0 contour characterising (prosodic) clause boundaries, which is different from the end of an utterance.

On the other hand, Shriberg et al. (1998) present the raw/normalised prosodic features, such as duration, pause, fundamental frequency (F0), energy and speaking rate, using a CART-style decision tree classifier. The authors confirm previous findings that prosodic features are not useful for all tasks of dialogue acts but work for some of them, such as question detection, incomplete utterance detection and agreements detection. A follow-up

study proposed by Stolcke et al. (2000) also modelled raw/normalised prosodic features using a decision tree. The classification model used was HMM-based generative, and the prosodic features identified were the duration, pause, pitch, energy and speaking rate. The reported accuracy on the Switchboard-DAMSL dataset (Core and Allen, 1997) was 38.9% using only prosodic features. When the reference transcripts were used along with the discourse contexts and n-grams as features, the obtained accuracy was 71%. In addition, when ASR outputs were used along with an integrations of multiple features – the prosodic, dialogue history and lexical features – the obtained accuracy was 65%.

In another study, conducted by Fernandez and Picard (2002), a normalised autocorrelation method with a Gaussian window was applied, to extract a set of pitch candidates. Dynamic programming was then used to select the best sequence of pitch values, by suitably defining an optimisation function that penalised large octave and voicing-to-invoicing transitions (Boersma, 1993). The final feature set included measurements related to pitch, energy, and duration. Although the study provided only preliminary results on the CallHome Spanish database (Finke et al., 1998), its findings show an improvement over previous works. The study concluded that SVMs provide promising results for future research on dialogue acts, particularly if lexical-based features are used as well. Similarly, Surendran and Levow (2006) explore the use of prosodic features alongside lexical features, using a combination of SVMs and HMM for dialogue act tagging. Results show that the addition of prosodic features improves the performance for some classes, namely instructions, acknowledgements, and all queries other than binary ones.

The investigations by both Shriberg et al. (1998) and Stolcke et al. (2000) are most similar to the work presented in this thesis, in using prosodic features for metadiscourse tagging. However, instead of using decision trees to model prosodic features, here raw continuous values of prosodic features were used and combined with other textual-based features. In addition, the focus for this study is to apply another source of information (*i.e.* prosodic features) along with other textual-based features, which can work for all classes, not just for particular ones.

5.3 SVM-based Metadiscourse Tagging

In this section, the SVM-based model for metadiscourse tagging is explored in more detail. First a more detailed discussion of the features used is presented. Then, the design of the SVM model is introduced.

5.3.1 Features

To build the MDT-SVM model, a variety of lexical, syntactic and prosodic features were designed. The aim was to capture linguistic properties associated with the metadiscourse tag expressions, as well as discourse properties associated with individual utterances. For the purposes of analysis, these features were partitioned into two groups: textual and prosodic features. Except for the prosodic features, all of the other features had count values indicating the frequency of that feature in the given utterance.

Textual Features

It is necessary to select the set of features based on how representative they are for the phenomena of metadiscourse. Recall the annotation experiments in Chapter 3; the participants were able to detect metadiscourse occurrences using only lecture transcriptions. In particular, they identified sets of words that signal the discourse function of the tag in question. This indicates that the most important features for the MDT-SVM model are those extracted from text materials. The same procedures were followed by Correia et al. (2014a) when developing the model to tag metadiscourse in TED Talks.

As noted in Chapter 3 (annotation), the manual transcriptions used for the annotation are not perfect, as they lack disfluencies that typically emerge in speech settings, such as filled pauses, fragments, repetitions or false starts (Moniz et al., 2012). These artefacts seem not have an effect on the overall comprehension of the text (Jones et al., 2003, 2005). In addition, their effectiveness has been proven in other tasks, such as automatic speech recognition (Stouten et al., 2006) and summarisation (Xiaodan and Gerald, 2006). However, they may limit the choices of using other language processing tools that are often used to extract other features (*e.g.* named entity recognition) as these models were trained on well-written texts with grammatical structures that differ from spoken materials, and may not cope with the errors in this type of transcription (Hayes et al., 1986). Hence, this will limit the focus further on lexical features, in particular word n-grams, and the use of other textual-based features will be as supporting ones. In summary, the list of features used for MDT-SVM model are as follows.

Words N-grams: As mentioned before, this is the main feature and includes all word n-grams in the given utterance. As with Correia (2013), an analysis study was conducted to further validate the appropriateness of this type of feature, by extracting the top n-grams (unigrams, bigrams and trigrams) from the annotated sentences, with respect to the four main metadiscourse categories, as shown in Table 5.1. It is necessary to clarify that the n-grams for the individual list include both disciplines, Physics and Economics. There is a clear difference

n-gram	Lectures	ML	DO	SA	IA
1	the, to, and, a, that, is, you, so, it, this	the, of, is, and, a, you, mean, called, define, sorry	the, to, of, we, is, talk, sum, you, about, going	the, to, of, a, is, this, example, very, going, important	the, you, to, say, think, so, guy's , might, but, going
2	in the, going to, of the, this is, and the, to the, if you, to be, want to, is a	that means, called the, which is, is the, I'm sorry, is called, it is, so that, it means, called a	of the, going to, in the, talk about want to, about the, last time, the following to sum, I will	going to, example of, want to, let's say , a very, that the, want to, of the, very important, and I	you guy's, going to, you might, you know, you hear, might think, you will, you can, you say
3	a lot of, we are going, are going to, going to be, I want to, is going to, I'm going to, you have to, you're going to, to talk about	is the same, which is the, is called the, is defined as, I'm going to, and that means, the definition of, I'm sorry I, that means we, in other words	I'm going to, to talk about, I want to, a little bit, in the following, last time we, I told you, come back to, in this course, last lecture we	I'm going to, is going to, I want to, you have got, an example of, is very important, the most important, in this course, one of the, very very interesting	and you guy's, you might think, you hear it, you might say, you guy's make you guy's are, you know what, I'm going to, most of you, do you think

TABLE 5.1: Top 3-grams in both Physics and Economics lectures, in contrast to the annotated four main metadiscourse categories.

between n-gram for the whole corpus and those for the annotated categories. For example, the *Metalinguistic comments* (ML) category ranks as top word n-grams that includes words such as 'mean', 'define', 'sorry'. These words are often used in the specific metadiscourse tags under the ML one. For instance, the word 'mean' can be used in expressions such as 'What I mean about this is', which indicates *Clarifying* (CLA). Another example is the use of the word 'sorry' in phrases such as 'I am sorry that's wrong', which corresponds to the metadiscourse tag of *Repairing* (REP).

In the *Discourse Organisation* (DO) category the top-ranked word n-grams contain words such as 'talk', 'told', 'last', 'sum'. Again, these phrases can be used in expressions such as 'Let's talk about'[1], 'I told you in the last lecture about', or 'In sum', which indicate the specific metadiscourse tags of *Introduction* (INT), *Reviewing* (REV), and *Conclusion* (CON), respectively. Similar observations were noticed when examining the most frequent n-grams in *Speech Acts*, where words include 'important', 'example', 'very', which can be used in a set of expressions that indicate *Emphasising* (EMP) and *Exemplifying* (EXE) specific tags. The *Interaction with Audience* includes 'guys', 'think', 'might', which can be used in specific tags such as *Managing Comprehension* (MAC) or *Anticipating the Audience Response* (AAR). In summary, it is clear that there are some words that are more representative for the task than others. In addition, there are also some general n-grams that are shared across categories, such as 'going to', and 'want to'.

There are additional two feature settings that need to be considered with regards to the above features: the inclusion of stop words and the use of word lemmas. Removing these was due to the fact that these words carry no meaning and hence there is no value to including them in the model (Osińsk and Weiss, 2005). It has been further reported that filtering

[1]These set of expressions were taken from the annotated dataset

them out would enhance the model performance in some applications, such as document indexing and retrieval, and topic modelling (Catarina and Bernardete, 2003, Osiński et al., 2004, Wang and McCallum, 2006). However, this is not the case in other applications, such as sentiment analysis (Lee and Ng, 2002, Maas et al., 2011, Paltoglou and Thelwall, 2010) where keeping them improved the results. Furthermore, it can be seen from Table 5.1 that *stop words* appear in every category. Hence, keeping them may be important for the task of metadiscourse tagging. The same observations were taken into account in Correia (2013) for the task of metadiscourse tagging in TED Talks. In addition, Correia (2013) noticed that removing stop words gave worse results than keeping them.

The other consideration to be taken into account is the *word lemma*,, which means grouping together with the same part of speech tag differently inflected forms of words that are syntactically different but semantically equal. This would reduce the number of features needed for the task. For example, the words 'sees' and 'saw' are grouped into the term 'see'. Including word lemmas along with other features significantly enhanced the model performance for the task of sentiment analysis, as reported by Mullen and Collier (2004). However, this is not the case for metadiscourse tagging in TED Talks, as the author reported that when the word lemma was combined with other features, it enhanced the classification model for some categories but not all Correia (2013). Hence, this has also been validated for the set of experiments presented in this chapter.

Part-of-speech Tags: The grammatical structure of the word n-grams might play an important role in classifying metadiscourse tags. Hence, the frequencies of part-of-speech tags (POS) of unigrams, bigrams and trigrams in the utterance are included. In total, there are 36 POS provided by the Stanford Parser[2] (Klein and Manning, 2003). This type of feature serves as a crude form of word sense disambiguation Wilks and Stevenson (1998): for example, it would distinguish the different usages of the word talk: 'Let's talk about' (indicating *Discourse Organisation*, in particular introducing a topic) versus 'This is a very important talk' (indicating *Speech Acts*, in particular emphasising).

Positional Information: A hypothesis in this work was that the position of an utterance in the lecture might make a difference in classifying metadiscourse tags. For example, the lecture might often begin with one of the *Discourse Organisation* tags such as Introducing topics, proceed with reviewing some concepts from the last lecture or common knowledge, and conclude by summarising the lecture content. As a rough approximation to determine this type of structure, each word was tagged according to whether it appeared in the first quarter, last quarter, or middle half of the lecture. In addition, the following features are included to the MDT-SVM model:

[2] http://nlp.stanford.edu/software/lex-parser.shtml

- Length of the utterance – how many words are in the utterance;
- Distance to the last occurrence – number of utterances between the current utterance and the last occurrence of the metadiscourse tag.

The inclusion of these three features was motivated by the analysis of two lectures during the pilot study, as described in Section 3.2. A substantial number of metadiscourse categories were frequent at the beginning of the lecture. This also occurs in consecutive utterances, as demonstrated in Tables 3.4 and 3.5, for Physics and Economics, respectively.

Prosodic Features

Prosody has been proved to carry important information and meanings related to discourse organisation and various information extractions tasks (Shriberg et al., 2000). In particular, pause duration (PD) and pitch (F0) serve as good indicators of the occurrences of metadiscourse tagging. To extract these prosodic features, the manual transcripts were aligned to the audio data, using a multi-genre broadcast media transcription system (Saz et al., 2015). More details on the alignment process are given in Section 4.3.4. In the following, a detailed extraction process of these feature is provided.

Pause duration (PD): This is defined, in this work, to be the time difference between the end time of a word and the start time of the immediately following word. The word boundaries were extracted by aligning the reference transcripts with the audio data, as mentioned above. These times were re-scaled in order to have a unitary range.

Pitch-based (F0): Pitch frequency mean and variance from the fundamental frequency (F0) were extracted using ESPS Entropic (1993), with the pitch tracking algorithm get F0 function, and sampled every 10 msec. Then, the words' temporal information provided by the alignment process was used to compute the mean and variance of F0 for each word. In addition, the mean F0 was further normalised against the lecturer's average pitch, in order to have a speaker independent feature.

5.3.2 Model

As a classification model, a linear SVM is used, as it is one of the most robust and successful classification algorithms, and is often used as a baseline performance in related tasks, such as classifying speech acts in email messages (Cohen et al., 2004, Qadir and Riloff, 2011), sentiment analysis (Wilson et al., 2009), and dialogue act tagging (Surendran and Levow, 2006). SVMs are binary classifiers, and are based on the idea of maximising the margin, *i.e.*

maximising the minimum distance from the separating hyperplane to the nearest example (Cortes and Vapnik, 1995).

Given a set of training sentences each with a d-dimensional vector, with labels yes or no (or -1 and 1), the weights of a linear SVMs can be learned $w \in \mathbb{R}^D$ with a threshold $b \in \mathbb{R}$ to predicate the class of new instance, as follows:

$$f(x) \;=\; sign(w^T x + b) \tag{5.1}$$

where the value of $f(x)$ is between $-\infty$ and $+\infty$, normalising the values becomes between 0 and 1 (Platt, 1999).

However, here an utterance can be labelled with one of a set of multiple metadiscourse tags. Thus, an extension to the basic SVM is used to handle the multiclass classification case. In particular, it fits one classifier per class and is handled according to an OVA approach, which means that for each classifier the class is fitted against all the other classes. This reduces the classification problem from classifying n classes into n binary problems, where each problem discriminates a given class from the other $n-1$ classes. Rifkin and Klautau (2004) states that this approach, although simple, has a level of performance that is comparable to more advanced modelling approaches.

More formally, the task of metadiscourse tagging is conducted as multiclass classification. The training dataset consists of utterances that belong to n different classes, and the goal is to construct a function to correctly predict the class of a new utterance. That is, n different binary classifiers were trained, in which f_i is the i^{th} classier to predict the class for new utterance x, as follows:

$$f(x) \;=\; argmax_i \; f_i(x), \tag{5.2}$$

where class i contains all the utterances from that particular class.

		Physics	Economics
	MD Tag	#	#
Metalinguistic	REP	82	91
	REF	180	64
	CLF	18	36
	CLA	276	215
	MAT	520	297
	Total	1076	703
Discourse Organisation	INT	182	296
	CON	96	115
	DEL	79	73
	COT	18	26
	ENU	441	465
	PHO	112	186
	REV	733	610
	PRE	447	355
	Total	2108	2126
Speech Acts	EMP	1075	926
	EXE	798	823
	ARG	39	9
	SUG	10	20
	Total	1922	1778
Audience	MAC	216	110
	AAR	107	42
	Total	323	152
	Overall	5429	4759

TABLE 5.2: A statistical summary of all the tags in the gold standard dataset for each discipline, after removing utterances that contains more than one tag.

5.4 Experiments and Results

5.4.1 Experimental Setup

Dataset

To validate the MDT-SVM model described in this chapter for the metadiscourse tagging task, the annotated metadiscourse dataset described in Chapter 3 was used. Since the MDT-SVM model described in this chapter is conducted as a multiclass classification task, the occurrences of the metadiscourse described in Table 3.7 are refined by removing those utterances that contain more than one class. Thus, each utterance could have one class at most. The final set of occurrences, which is used for training and testing, is presented in Table 5.2. In total, 1278 utterances out of 11,466 have multiple metadiscourse tags across disciplines. This constitutes 11.1% of the total annotated utterances. Removing those left

a corpus of 10,188 utterances that have only one occurrence of metadiscourse tags across disciplines.

Adding Annotations to ASR Outputs

An important issue in evaluating the metadiscourse tagging model over ASR system output is how to obtain the system annotations. In general, the metadiscourse tagging model should not be penalised for the errors in the transcriptions or classifying words as part of metadiscourse that were not actually said in the lecture. ASR outputs are considered distinct transcriptions that need to be annotated by a human and by the system. However, human annotation of ASR outputs is not feasible, as some parts in the transcriptions do not make sense. This means evaluating the model performance in the case of ASR outputs should be the same procedure as in the reference transcriptions.

In this regard, this study followed Burger et al. (1998), Galibert et al. (2011), and Hirschman et al. (1999), who proposed a method that projects the clean reference on to the noisy text, in order to build a new reference. The new reference then allows the application of the clean text methodology. However, their objectives differed from this study, as in their case a sentence could have more than one tag. Hence, finding the word sequence suitable for this tag was important for their task. In the case of the metadiscourse tagging task, the only interest in aligning reference transcriptions with ASR outputs is at sentence-level, as each sentence (or utterance) has only one metadiscourse tag. The steps taken for this process are as follows:

1. Use the timestamped reference transcription that results from the alignment process, as described in the previous chapter, in Section 4.3.4.
2. Extract the start and end times of each sentence in the reference transcriptions that contain the annotations.
3. Select a tolerance time interval.
4. Find the ASR words within the tolerance intervals.
5. Keep track of the corresponding metadiscourse tag of the sentence in the reference.
6. Assign the metadiscourse tags to these sets of words.

Evaluation

The task was evaluated using a stratified 10-fold cross-validation (CV), where each fold contains approximately the same number of class samples. As a metric, the results reported

are computed as the average of the 10-fold CVs, using the standard classification metrics precision, recall and F measure (Rijsbergen, 1979), which will allow evaluation of the level of difficulty of classifying each metadiscourse tag. They are the recommended metrics when dealing with highly skewed datasets (Davis and Goadrich, 2006).

$$Precision = \frac{TP}{TP + FP} \tag{5.3}$$

$$Recall = \frac{TP}{TP + FN} \tag{5.4}$$

$$F = 2 \times \frac{Precision \times Recall}{Precision + Recall}, \tag{5.5}$$

where T stands for true, F for false, P for positive and N for negative. However, there are two conventional methods of computing the precision and recall in a multiclass classification problem: *micro-averaging* and *macro-averaging*. While the former is calculated by constructing a global contingency table and then computing precision and recall using these sums, the latter is calculated by first calculating precision and recall for each category and then taking the average of these (Davis and Goadrich, 2006). In this thesis the *macro-averaging* approach is used.

$$Precision_{macro} = \frac{1}{C} \sum_{i=1}^{C} \frac{TP_i}{TP_i + FP_i} \tag{5.6}$$

$$Recall_{macro} = \frac{1}{C} \sum_{i=1}^{C} \frac{TP_i}{TP_i + FN_i} \tag{5.7}$$

Tools

The tool used to conduct the set of experiments presented in this chapter, including the classifier, is scikit–learn Pedregosa et al. (2011) which is based on LIBLINEAR algorithms (Fan et al., 2008). This is in addition to the NLTK toolkit, which is used to prepare the features set, such as word n-grams and lemmas. For the English word list the one provided by this tool was used. To formulate the dictionary for the n-grams, the most frequent words in the dataset were kept. Other approaches may be based on the information gain observed when keeping a certain n-gram, such as the work presented by Correia et al. (2014a). However, they used the presence of the feature in the sentence instead of the frequency,

Feature	Physics			Economics			Overall		
	P	R	F	P	R	F	P	R	F
LEX-TGM	17.39	15.46	16.37	19.79	16.84	18.20	18.59	16.15	17.29
LEM-TGM	25.99	21.74	23.68	27.54	23.68	25.46	26.77	22.71	24.57
POS-TGM	11.31	10.38	10.83	14.82	12.01	13.27	13.07	11.20	12.05

TABLE 5.3: Results of the decision tree model using *tri*–grams of words (LEX), Lemma (LEM) and POS tags for all metadiscourse tags.

whereas in this thesis the frequency is taken into account when selecting the word list. The Stanford Parser[3] was used to tag each sentence with POS tags (Klein and Manning, 2003).

Experiments Settings

It is important to note that all the results reported here are for the metadiscourse tagging using specific tags. Results for generic metadiscourse tags classification are provided in section 5.4.7. In addition, in the following sections, the results were reported in terms of Precision, Recall and F measure for each discipline individually. Consideration was taken to test the effects of using both a combined model and splitting each discipline. In the following sections a distinction is made in terms of performance between the two disciplines. As mentioned previously, considerations were made to include or exclude stop words from the presented set of experiments. However, excluding stop words decreased the classification performance when n-grams features only were used, in the order of 4.3% in the F-measure score. Hence, for the rest of the experiments presented in the following sections, stop words were included.

5.4.2 Preliminary Experiments

To make an initial estimate about the metadiscourse tagging task and the data, a rule-based system, specifically a decision tree (DT), was used before experimenting on the metadiscourse tagging task using more advanced models, such as SVM. These preliminary experiments with DTs approximately replicated previous work by Correia et al. (2014a), which enabled a comparison with the TED talks corpus. A DT classifier consists of a set of rules that are learned in the training process. In this specific setup, the Weka[4] implementation of the C4.5 algorithm (J48) was used. In addition, the same preprocessing setup of the data used for the SVM model was followed.

To conduct the preliminary experiments, a variety of lexical-based features were used: trigrams of words, lemmas and POS. These features were used for the DT baseline because

[3] http://nlp.stanford.edu/software/lex-parser.shtml
[4] http://www.cs.waikato.ac.nz/ml/weka/

	Physics			Economics			Overall		
Tag	P	R	F	P	R	F	P	R	F
INT	37.50	28.02	32.07	51.60	48.99	50.26	44.55	38.51	41.17
CON	30.56	22.92	26.19	33.89	34.78	34.33	32.23	28.85	30.26
EMP	52.51	47.63	49.95	54.97	50.76	52.78	53.74	49.19	51.37
EXE	66.45	62.78	64.56	73.21	60.75	66.40	69.83	61.77	65.48

TABLE 5.4: Results of four metadiscourse categories: Introduction (INT), Conclusion (CON), Emphasising (EMP) and Exemplifying (EXE) using Decision tree mode with trigram of lemma as features.

they were the main features used in later experiments with the MDT-SVM model (see Section 5.4.3). They were also used in previous work on TED talks (Correia et al., 2014a) with the DT model, which facilitated the replication process on the OCW dataset. Table 5.3 shows that the trigram of lemma proved to be representative for the metadiscourse tagging task using DT in both disciplines. Using word trigrams yielded poor results because many of the features in the set were not sufficiently distinct for the task using the DT model. However, some of the results matched the observations seen with the MDT-SVM model, such as the case with the POS tags as features, which gave a poorer performance compared with the other features (see Section 5.4.3). This observation was also noticed with previous work on TED talks (Correia et al., 2014a).

To compare the DT experiments on the OCW dataset with previous work on TED talks, the model was applied to the four specific metadiscourse tags as studied by Correia et al. (2014a): Introduction (INT), Conclusion (CON), Exemplifying (EXE) and Emphasising (EMP). Table 5.4 shows the results of the four categories using the DT model; the trigram of lemma yielded the best results as indicated previously. In general, the results followed the normal expectation that there was a correlation between the frequency of the tag and its classification performance. For example, a less frequently occurring tag had a poorer performance compared with other tags, such as with CON, especially in Physics. However, this was not the case for EMP, which was the most frequently used tag in the dataset compared with EXE, although EXE had a better performance. These observations agree with the previous work studied by Correia et al. (2014a) and with our observations using the MDT-SVM model, discussed later in this chapter. In addition, the general performance of the DT model on the OCW dataset was much lower than the performance reported on the TED talks (Correia et al., 2014a). This highlights the difficulty in detecting the metadiscourse tags in academic lecture datasets.

In summary, the performance of the preliminary experiments for the task of metadiscourse tagging on the OCW dataset needs further improvements, especially if the aim is to use the automatically detected tags for downstream applications, such as for thematic discourse

Feature	Physics			Economics			Overall		
	P	R	F	P	R	F	P	R	F
					POS				
UGM	18.11	07.97	11.07	19.72	08.16	11.54	18.92	08.07	11.31
BGM	28.30	11.00	15.84	28.64	11.10	16.00	28.47	11.05	15.92
TGM	26.48	14.19	18.48†	28.33	15.68	20.19†	27.41	14.94	19.34
					LEM				
UGM	46.70	28.85	35.67	52.26	34.65	41.67	49.48	31.75	38.67
BGM	46.93	32.28	38.25▽	48.61	33.21	39.46	47.77	32.75	38.86
TGM	47.67	33.18	39.13‡ ∗	52.13	38.27	44.14‡ ∗	49.90	35.73	41.64
					LEX				
UGM	47.12	29.28	36.12	52.18	34.50	41.54	49.65	31.89	38.83
BGM	47.38	32.79	38.76◇	51.70	37.61	43.54◇	49.54	35.20	41.15
TGM	47.76	33.14	**39.13**	52.22	38.30	**44.19**	49.99	35.72	41.66

TABLE 5.5: Results of using n-grams frequencies of words, lemma and POS tags. LEX denotes word n-grams, LEM refers to lemma n-grams. † denotes statistically significant results when compared to the best results within the POS features experiments. ‡ indicates statistically significant results and ▽ denotes insignificant difference when compared to the best results within the LEM features experiments. ◇ denotes insignificant difference when compared to the best results within the LEX features experiments. Bold face denotes significant results within LEX features experiments and overall. ∗ denotes insignificant difference when compared with the LEX-TGM features.

segmentation. Thus, a more advanced model, such as SVM, is required for the tagging task. Another advantage of using the SVM model is that it allows both discrete and continuous features such as prosodic features, to be easily integrated. This proved effective in similar tagging tasks, such as dialogue acts (Surendran and Levow, 2006).

5.4.3 Feature Combinations

In this section, the results of different features and feature combinations using the MDT-SVM model are reported. For the purposes of analysis, these features are partitioned into three groups: n-grams of *Word, Lemma and POS*, *Positioning Length* and *Prosodic Cues*. In addition, the effects of using ASR outputs are also reported for some of the feature combinations that reported best on the reference transcriptions.

It is important to note that the significance test is perfumed by evaluating each experiment using 50-fold cross-validations and then computing the F1-scores. In particular, a t-test was used to check the statistical significance of the result of each experiment compared to the best obtained results. A t-test is usually used to compare the means of two groups if they are significantly different from each other (Zimmerman, 1997). For instance, in Table 5.5 the best result is obtained by using lexical tri-gram (LEX+TGM) features, as indicated by bold-face. Then the set of experiments that give the best results in other features, such as POS+TGM or LEM+TGM, are compared to the best result across all features (*i.e.* LEX+TGM).

Feature	Physics			Economics			Overall		
	P	R	F	P	R	F	P	R	F
LEX*	47.76	33.14	39.13	52.22	38.30	44.19	49.99	35.72	41.66
LEX+LEM	46.21	34.15	39.28	51.26	39.56	44.66 *	48.74	36.86	41.97
LEX+POS	45.12	36.14	40.13	47.83	40.65	43.95	46.48	38.39	42.04
LEM+POS	44.42	36.18	39.88	44.63	37.3	40.64	44.53	36.74	40.26
LEX+LEM+POS	46.66	39.32	**42.68**	48.00	42.43	**45.04**	47.33	40.88	43.86

TABLE 5.6: Results of using a combination of n-grams of words (LEX), Lemma (LEM) and POS tags, simply (POS). Bold face denotes significant results and * denotes insignificant difference.

N-grams of Word, Lemma and POS

The first experiment settings tested were the use of n-grams of words, lemmas and POS. Unigram, bigram, trigram and a combination of these were tried. It is important to note that the bigram features include unigram features, and the trigram features include both unigrams and bigrams. Table 5.5 reports the results for each pair of features/discipline.

In general, the results show that the use of syntactic features only decreases the model performance compared to other n-grams features used in all disciplines. Also, the use of word n-grams provides the most significant results (average F1-score 41.66%) in both disciplines. Results also show that out of all textual n-grams features (*e.g.* POS, or lemma or words), the use of trigram features provides the most significant results compared to unigrams and bigrams; this observation was consistent in both disciplines. It is also noticeable that the use of lemma and words trigrams have approximately similar performance in both disciplines; the difference in performance is insignificant, as indicated by * in Table 5.5. For example, for Physics lectures the model provides the same F1-scores results, 39.13%. Similarly, for Economics lecturers the results were 41.14% and 41.19% when lemma and words trigrams are used, respectively.

The results of previous experiments were inconclusive regarding the use of n-grams features to classify metadiscourse tags. Further investigation is needed to gain further insight into the previous results having roughly similar results when either word or lemma features were used. In particular, it is crucial to know whether this similarity is due to the fact that these two features represent the same information, or because they complement each other. It would also be interesting to know whether inclusion of the syntactic features would add any value to these lexical combinations. To test these assumptions, Table 5.6 shows the results of the experiments of a combination of the trigrams of words, lemmas and POS tags. The combination of all three of these features significantly improved the overall results of the MDT-SVM model, to 43.86%, compared to 41.66% and 41.64% when using only the trigrams words and lemma, respectively, as shown in Table 5.5. Another important consideration is the difference between the two disciplines, since classifying metadiscourse using Economics

Feature	Physics			Economics			Overall		
	P	R	F	P	R	F	P	R	F
LEX+LEM+POS	46.66	39.32	**42.68**	48.00	42.43	45.04*	47.33	40.88	43.86
LEX+LEM+POS+Length	47.58	34.02	39.67	52.79	40.06	**45.55**	50.19	37.04	42.61
LEX+LEM+POS+Position	44.74	37.42	40.75	49.89	39.62	44.17	47.32	38.52	42.46
LEX+LEM+POS+Distance	31.69	29.84	30.74	47.65	33.88	39.60	39.67	31.86	35.17
LEX+LEM+POS+Length +Position+Distance	43.77	25.04	31.86	45.42	33.32	38.44	44.59	29.18	35.15

TABLE 5.7: Results of using positional information (Length, Position, and Distance), along with other features including lexical (LEX), lemma (LEM), and Part-of-Speech Tags (POS). Bold face denotes significant results and * denotes insignificant difference.

lectures provides far better results than Physics lectures. For instance, in the settings of the best combination of features (n-grams of words, lemmas and POS tags) the overall F score of Physics lectures was 42.68%, compared to 45.04% in Economics lectures. This is despite the fact that the total number of metadiscourse tag occurrences in Physics is higher than those in Economics. This may indicate that the expressions used in Economics lectures are less variable than those in Physics lectures.

Positional, Length and Distance

In this section, experiments conducted using features that exhibit some of the discourse structure are reported. These features are: the length of the sentence, the position of the sentence in the lecture, and the distance between the current sentence under classification and the last occurrence of a metadiscourse tag.

Table 5.7 shows the results of using these positional features individually and also when combined with the best combination of n-grams of words, lemmas and POS tags from the previous section. Results indicate that most of the positional features have no impact on the classification performance. However, among the aforementioned features the length feature achieved the best results, particularly for Economics lectures. For instance, in Economics lectures the F score, when adding the length information over the previous n-grams features, increases to 45.55%; but this improvement is not statistically significant. Similarly, for Physics lectures the overall results decreased: from 42.68% when only the n-grams of words, lemmas and POS were used, to 39.67% when adding the length information. The performances of using the rest of these features, namely position, distance and combinations of all, are not significant. In general, the small improvement in the performance from using such features may indicate that these types of features cannot be generalised as much as the n-grams features for the metadiscourse tagging task.

Feature	Physics			Economics			Overall		
	P	R	F	P	R	F	P	R	F
LEX+LEM+POS	46.66	39.32	42.68	48.00	42.43	45.04	47.33	40.88	43.86
LEX+POS+F0	46.64	40.92	43.59	47.90	43.62	45.66	47.27	42.27	44.63
LEX+POS+PD	46.16	41.35	43.62	49.28	45.17	47.14	47.72	43.26	45.38
LEX+POS+F0+PD	46.09	42.25	**44.09**	50.31	47.36	**48.79**	48.20	44.81	46.44

TABLE 5.8: Results for adding prosodic features (F0, PD) to reference transcriptions. Bold face denotes significant results.

Prosodic Cues

The final set of experiments considered the inclusion of the prosodic cues. In particular, pitch-based features and pause duration were used. Table 5.8 presents the results of these experiments, first individually, then combined. Pause duration was found to have a better influence on the results than using F0, and this is consistent in both disciplines. In Physics lectures the improvement was from 42.68% to 43.62% in F score. Similarly, the F score increased from 45.04% to 47.14% for Economics lectures. This can be attributed to the fact that pause duration can capture boundary information between words, and this may serve as an indication of metadiscourse instances. For example, lecturers often tend to pause after saying something important (EMP tag) or even when they introduced the topic of the lecture (INT). The purpose here is to allow the students to absorb the information just given. This can be true for most of the metadiscourse tags, as the expressions used to signal the functions of each of these tags have a main purpose: to engage the students during the lecture. In addition, the combination of prosodic features seems to be statistically significant in both disciplines, with an overall F score of 46.44%. In general, the inclusion of prosodic features was found to have more impact compared to positional and length features for the task of metadiscourse tagging.

5.4.4 Comparison to a Naive Baseline

To further explain the performance of the MDT-SVM model, the best results were compared with a very naive baseline. These results were obtained by the combination of lexical n-grams and prosodic features using the MDT-SVM model (see Table 5.8). In the naive baseline system, all utterances were labelled as the most frequent metadiscourse tag in the OCW dataset, which is EMP as indicated in Table 5.2. The F1-scores were then computed for both disciplines. The main intuition of this experiment was to see how well the results of the best systems of the MDT-SVM model compared with the results from using a totally wrong system for the task of metadiscourse tagging. The naive baseline system yielded F1-scores of 3.27 and 3.01 for Physics and Economics, respectively. In addition, as expected, the best

Feature	Physics			Economics			Overall		
	P	R	F	P	R	F	P	R	F
LEX	34.52	20.66	25.85	39.95	25.08	30.81	37.24	22.87	28.33
LEM	33.90	20.44	25.50	39.33	24.73	30.37	36.62	22.59	27.94
LEX+POS	34.08	20.34	25.48	39.60	24.69	30.42	36.84	22.52	27.95
LEX+PRO	36.05	26.27	30.39*	34.90	29.13	31.76	35.48	27.70	31.08
LEX+POS+PRO	34.49	28.76	**31.37**	35.66	32.34	**33.92**	35.08	30.55	32.65

TABLE 5.9: Results of features combination on ASR transcriptions. Bold face denotes significant results and * denotes insignificant differences.

Model	Physics			Economics			Overall		
	P	R	F	P	R	F	P	R	F
In-disciplines	46.09	42.25	44.09	50.31	47.36	48.79	48.20	44.81	**46.44**
All-disciplines	-	-	-	-	-	-	41.32	39.60	40.44

TABLE 5.10: Results show a comparison of in-discipline and all-domain metadiscourse classifications.

results obtained using the MDT-SVM model significantly outperformed this naive baseline, which provided very poor results in both disciplines.

5.4.5 Effects of using ASR Outputs

To investigate the effects of metadiscourse classification on ASR outputs, a set of experiments was conducted using the ASR outputs which have around 28% WER in both disciplines, as described in Section 4.3. Table 5.9 shows the best results for different feature combinations on ASR outputs that were reported as previously, on reference transcriptions. It is important to note that the features used in these experiments are extracted from ASR transcriptions. In general, as expected, the classification performance degraded when ASR transcriptions were used. In addition, in both disciplines the use of word n-grams gave better overall results (average $F1$-score is 28.33%) than the use of lemmas (27.94%). However, the inclusion of POS tags degraded the classification performance further compared to the addition of prosodic features. For example, the overall $F1$-score in both disciplines is 27.94% when POS tags used. However, the effects of adding the prosodic cues was significant, with an average $F1$-score of 31.08%. This can be attributed to the fact that the POS tagger used was trained on cleaner setups, such as written text. It seems that combining all knowledge sources achieved more improvements in model performance, with an overall $F1$-score of 32.65%. Similarly, the case with reference transcription is that the model performance across different features was better for Economics lectures.

	Physics			Economics			Average		
Tag	P	R	F	P	R	F	P	R	F
ML	60.43	60.94	60.68	49.21	49.64	49.42	54.82	55.29	55.05
DO	51.62	47.85	49.66	60.97	58.72	59.82	56.29	53.29	54.74
SA	60.61	59.04	59.81	66.86	67.02	66.94	63.74	63.03	63.38
IA	48.91	40.36	44.23	67.11	56.04	61.08	58.01	48.20	52.66

TABLE 5.11: Results for generic tags metadiscourse classifications.

5.4.6 In-domain vs. All-domain Classifications

Before conducting the set of experiments presented in this chapter, there is one decision that needed to be taken with respect to modelling strategy. One could either combine disciplines and develop one model for all, or show the results for each discipline individually, and then report the results of the average while highlighting the differences between disciplines. Since part of the overall aim of this work is to shed light on per-discipline classification in these courses of experiments, and to address the differences between the disciplines in terms of performance, feature combination and tag granularity (generic vs. specific), it was decided to keep a distinction between the two disciplines in terms of the reported results throughout the set of experiments conducted.

In addition, to further support the hypothesis that in-domain classification works better than when combining the two disciplines, experiments were conducted that illustrate the difference in terms of performance in these two cases. The set of features reporting the best in previous settings of a combination of LEX and POS and PRO was selected, namely the n-grams of words and POS tags and also the prosodic cues. The results presented in Table 5.10 confirm the hypothesis that the metadiscourse tagging task is domain-dependent, since making a distinction between disciplines gave significantly better results (the F1-score on average 46.44%) than when combining them (40.44%).

5.4.7 Generic vs. Specific Tags Classifications

In this section, per-tag scores are analysed at two levels: specific and generic. The purpose of this is to show the model performance across the different metadiscourse tags and highlight any variation between them. For this set of experiments, the best feature combinations found previously were used. These features are the word n-grams, POS tags and prosodic cues. It is important to note that the NONE tag is included in the classification process, which captures the majority of utterances that have no metadiscourse tags. This is because the NONE tag also provides very good results, often over 90% for the Precisions and 80% for the Recall in both disciplines in either specific or generic tags case. For this reason, in

		Physics			Economics			Average		
	Tag	P	R	F	P	R	F	P	R	F
ML	REP	75.61	79.49	77.50	70.00	71.59	70.79	72.81	75.54	74.15
	REF	91.67	85.94	88.71	85.48	75.71	80.30	88.58	80.83	84.51
	CLF	16.67	08.57	11.32	13.89	07.46	09.71	15.28	08.02	10.52
	CLA	45.26	45.09	45.17	66.82	67.14	66.98	56.04	56.12	56.76
	MAT	42.47	39.45	40.90	27.34	24.38	25.77	34.91	31.92	33.34
DO	INT	34.81	36.00	35.39	56.61	58.60	57.59	45.71	47.30	46.49
	CON	19.15	16.67	17.82	38.94	27.33	32.12	29.05	22.00	24.97
	DEL	23.38	23.08	23.23	38.03	36.49	37.24	30.71	29.79	30.24
	COT	11.76	09.09	10.26	20.00	25.00	22.22	15.88	17.05	16.24
	ENU	30.45	28.45	29.42	43.63	40.89	42.22	37.04	34.67	35.82
	PHO	40.97	38.56	39.73	58.41	61.11	59.73	49.69	49.84	49.73
	REV	58.54	56.95	57.73	64.74	66.43	65.57	61.64	61.69	61.65
	PRE	56.36	41.89	48.06	58.76	51.49	54.88	57.56	46.69	51.47
SA	EMP	50.61	46.94	48.71	56.55	53.38	54.92	53.58	50.16	51.82
	EXE	74.20	75.95	75.06	80.48	78.89	79.68	77.34	77.42	77.37
	ARG	61.54	48.00	53.93	00.00	00.00	00.00	30.77	24.00	26.97
	SUG	10.00	03.45	05.13	27.78	22.73	25.00	18.89	13.09	15.07
IA	MAC	55.35	40.89	47.04	83.64	67.65	74.80	69.49	54.27	60.92
	AAR	31.13	27.27	29.07	21.43	16.36	18.56	26.28	21.82	23.82

TABLE 5.12: Results for specific tags metadiscourse classifications.

the following the discussion about NONE is omitted, and the analysis is mainly about the metadiscourse tags inclusive.

Table 5.11 presents the results of the four generic tags, namely *Metalinguistics Comments* (ML), *Discourse Organisation* (DO), *Speech Acts* (SA) and *Interaction with Audience* (IA). In general and across both disciplines, it seems there is a correlation between having high occurrences and good classification scores. For example, SA tags have high occurrence and yields a high $F1$-score of 62.5%. However, this is not the case for some of the generic tags; for example, ML has a better $F1$-score (55.05%) than DO (54.74%), despite the fact that the occurrences of ML are fewer than DO, as indicated in Table 5.2. This can be attributed to the fact that the set of metadiscourse expressions in ML may be less variable than those in the DO tag. Moreover, the number of specific tags being mapped to one generic tag of DO is 8, compared to only 5 in ML. In addition, the analysis per-discipline suggests that generally the Economics lectures have better results than Physics lectures. This is consistent across all four tags, as also observed in the previous experiments.

Summary of Trends

For specific tags, Table 5.12 shows per-class scores for all the 19 tags in the metadiscourse scheme. For clarity, the scores are grouped by the general tag that these specific tags belong to. In general, the main observation is that many results are in line with normal expectations that more frequent tags have higher F1-scores. However, this is not the case for other tags, something which needs further explanation and analysis. The following analysis is focused on these cases and is organised according to the generic tags.

Metalinguistics Comments (MC): Among this set of tags, both REP and REF give the best result. This is despite the fact that they have low occurrences compared to both CLA and MAT, which occur in both disciplines a total of 491 and 817 times, respectively (see Table 5.2). This confirms the previous observation that getting high performance is not related only to having a high occurrence, but may also be related to the variation in metadiscourse expressions as well.

Discourse Organisation (DO): There are some tags in this set that have different observations than the normal expectations mentioned above. For instance, the PHO tag has low frequency of occurrence in both disciplines, but the performance (average F1-score = 49.73%) is better than the other tags in the DO set that have higher frequencies, such as INT. This can be attributed to the fact that the set of expressions that are usually used in referring to the PHO tag may be limited. Hence, the model is able to perform accurately compared to others. Similar observations have been also noticed with the DEL tag, especially in the discipline of Economics. Despite the fact that DEL has lower frequency of occurrence than CON, the classification performance is better, with an average F1-score of 30.24% compared to 24.97% for CON.

Speech Acts (SA): There are unusual observation for some of the tags among this set, for example with EMP and ARG. EMP has far more occurrences than the other tags, especially when compared with the ARG tag, as indicated in Table 5.2, but its performance is much lower than ARG in Physics lectures. This indicates that in such lecture sets the set of expressions used to refer to important points exhibits some variations. Hence, the model may need more example data to perform well.

Interaction with Audience: It has been noticed that there is inconsistency in the performance for some of the tags in this set across both disciplines, which can again be attributed to the variants in metadiscourse expression. For example, the MAC category has higher frequency of occurrence in Physics (216) than in Economics (110), but the F1-score in the latter is higher (74.80%) than the former (only 47.04%).

5.4.8 Discussion

Unfortunately, no direct comparisons can be made with previous work. This is because of the novelty of the annotated dataset used, as its been developed specifically to target metadiscourse in academic lectures as demonstrated in Chapter 3. Related previous work (Correia et al., 2015, 2014b, 2016) has developed a different dataset of metadiscourse and targeted a different application (building a tool to instruct students on how to do presentations) than the one considered in this thesis (segmentation and structuring academic lectures courses). Nevertheless, we tried to adopt their model in our newly annotated dataset in order to have a sense on how different the OCW dataset is in detecting and classifying metadiscourse compared with the TED talks corpus. More precisely, a decision tree model was developed in a similar fashion as the one used by Correia et al. (2015), and applied to only four metadiscourse tags, namely introduction, conclusion, emphasising, and exemplifying, as demonstrated in Section 5.4.2.

In general, the SVM model is able to classify metadiscourse tags at the two levels of granularity: generic and specific. The focus of this chapter was on specific tags, as they clearly reflect the functionality of each sentence in the lecture. The performance of generic tags is expected to be better than specific ones in both disciplines. This can be attributed to the high number of occurrences of the four generic tags compare to the specific ones.

In addition, experiments with features on specific tags show that the use of lexical-based features such as word and lemma n-grams is most important, and all other features can be used to improve the results further. Examples of these supporting features are the number of occurrences of POS tags, and the prosodic cues, particularly PD. These observations were noticed for both disciplines, Physics and Economics. It was also observed when ASR outputs were used instead of reference transcriptions. Nevertheless, when ASR outputs are used, the improvements achieved from using prosodic features are better than with the use of POS tags. This can be attributed to the fact that the utterance structure in ASR outputs has been lost due to the typical errors that arise in these types of transcriptions, such as insertions, deletions and substitutions. This would have a negative impact on the POS tagger.

Moreover, despite the facts that Physics lectures have higher occurrences than Economics lectures, the performance in the latter is better than in the former. This may be related to the fact that the Physics lecturer used a far wider set of expressions to perform the same metadiscourse functions. In Economics, meanwhile, its seems there is a relatively fixed set of expressions used throughout the lecture course, which explains the high performance in this discipline. Moreover, experiments on domain classifications indicate that the best classification performance is achieved when in-domain classification models are used, as illustrated in Section 5.4.6. However, when all disciplines are combined, as the results show, there are

no further improvements over the in-domain case. These conclusions are reasonable, as the model in the in-domain case will be more specialised to this domain, and adding different domains causes a degradation in performance.

All in all, the overall performance across specific tags shows that there is room for improvement. With the sparsity problems that comes from the fact that there are many different metadiscourse expressions that indicate the same functions. In fact, the reliance on lexical features such as word or lemma n-grams with such small dataset causes the sparsity problems. There is a need to either enrich the MDT-SVM model with more annotated examples of each metadiscourse tag, or develop a model with feature sets that are able to overcome the sparsity problem with such small numbers of annotated examples for each tag.

5.5 Conclusion

This chapter has presented the first model (MDT-SVM) to metadiscourse tagging using a combination of hand-engineered features and SVMs, using the developed corpus of metadiscourse tags, annotating at utterance-level on both manual and ASR transcriptions, as described in the previous two chapters (Chapter 3 and Chapter 4). The most effective features are a combination of n-grams of words, lemmas and POS tags, as well as the extracted prosodic cues. The results show that domain knowledge has an effect on the proposed approach. Hence, a distinction was made between the two disciplines throughout the rest of the experiments. In addition, as expected, the model performs well on higher-level metadiscourse tags (generic tags) and poorly on lower-level (specific) tags, in particular for low-occurrence tags such as *Arguing* (ARG). This can be attributed to sparsity problems in the datasets.

The purpose of this approach was to investigate the appropriateness of combining both high-dimensional textual features and low-dimensional ones, such as prosodic cues. This was in addition to developing the baseline model using the SVMs classifier. However, the approach suffers from sparsity problems and the model will not be able to generalise well for unseen n-grams. This limits its ability to solve the variants issue in metadiscourse expressions, thus limiting the effectiveness of this method. Continuous Bag-of-Words (CBOW) provides a promising property, as it can capture both the syntactic and semantic similarities between words, and thus solve the generalisation issue. A downside of the CBOW is that it ignores the word order completely, which is very important to retain when classifying metadiscourse tags. The next chapter attempts to solve this issue by exploiting both CBOW and CNNs.

Chapter 6

Improving Metadiscourse Tagging with CNNs

This chapter presents a study on the use of convolutional neural networks (CNNs) to alleviate the shortcomings of the approach of the MDT-SVM model discussed in the previous chapter. The set of experiments conducted here presents the best practice for configuring the CNN model for the task of metadiscourse tagging, particularly with regard to the use of word embedding and feature representation. Results show that CNNs outperform SVMs by a large margin, which proves the effectiveness of the method for the task. In addition, an in-depth analysis per class score shows some behavioural similarities of CNNs to SVMs, such as the correlation between high frequency and high performance. However, this is not the case for some of the tags, such as *Clarifying* (CLA). This observation leads to the need to conduct a further analysis to investigate areas of improvement for the classification models. The analytical study presents some suggestions for future work regarding the annotation scheme.

The chapter is structured as follows. Section 6.1 introduces the topic of this chapter, combined with the motivation and contributions of the presented work. Section 6.2 gives an overview of the basic structure of CNNs from an NLP prospective and how the features are represented in such a network, in addition to a review of previous work related to sentence classifications using CNNs. Section 6.3 presents the implemented CNN architecture. A summary of the results obtained from the set of experiments is given in Section 6.4, along with a concluding discussion and analysis.

6.1 Introduction

Neural networks (NNs) are powerful learning models that achieve state-of-the-art results in a wide range of supervised and unsupervised machine learning tasks in several research areas, such as speech recognition, image and video processing as well as natural language processing. In NLP, NNs are used for several tasks related to metadiscourse, such as sentiment analysis, discourse analysis and language modelling. The key for these models is to have continuous feature functions that are representative of the task. The popularity of NNs in such tasks comes from their ability to produce generic vectors for words or phrases by predicting the contexts in which the word or phrase occurs. This is also extended to obtain vectors that can represent whole paragraphs or documents. In this way, the word vectors are more powerful because statistical strength is shared with other vectors that have similar semantic and syntactic structures. This limits the need to use feature engineering approach that, although significantly successful in the past, does not consider this factor.

Two NN models are widely used in NLP tasks, namely CNNs (Lecun and Bengio, 1995) and recurrent neural networks (Elman, 1990). Convolutional neural networks can encode the features of a sequence of words into a fixed-size vector with short propagation paths, but they do not capture the grammatical structure of the sentence (Goldberg, 2015, Mou et al., 2015). While RNNs can encode sentence structural information by recursive semantic composition, they require long propagation paths (Erhan et al., 2009). For the task of metadiscourse tagging, finding those sets of words that are most representative for the task in the sentence and ignoring the rest of the words is important. Thus, CNNs are more suitable because they are designed to identify the most inductive features locally, regardless of their positions in the sentence, for the classification task at hand (Goldberg, 2015).

A key aspect of using CNN models or even any NN model for that matter is to represent the features with the use of dense, low-dimensional vectors instead of sparse, high-dimensional vectors. Note that these models only deal with core features, such as words and POS tags, and not their combination, in which each core feature is actually embedded into a d-dimensional space. One advantage of this approach is the limited need for the feature engineering stage, which is often required in linear models. Another advantage of embedding vectors is their generalisation power, because similar words have similar vectors. These embedding vectors can then be trained like the other parameters of the network with the use of several training algorithms, such as stochastic gradient descent (SGD). One disadvantage, however, is that it does require large amounts of data to train the network and thus produce the final set of vectors that can then be tuned for specific tasks through supervised training (Kalchbrenner et al., 2014). Other features, such as POS tags, can be treated in a similar fashion and hence produce dense vectors instead of sparse ones (Collobert et al., 2011). However, most

of the classification tasks, including metadiscourse tagging, have limited amounts of labelled training data available.

6.1.1 Motivations

Recently, various approaches have been proposed to generate continuous feature vectors without the need for labelled data for the task at hand. These methods use auxiliary tasks, such as a language model that is trained with the use of unlabelled data to produce feature vectors for the classification task. Another approach is the use of pre-trained word vectors that do not require any data, either labelled or unlabelled. For example, Kim (2014) proposed a simple CNN model that utilises existing pre-trained word vectors, such as Google word vectors (*word2vec*; Mikolov et al. (2013)). Despite its simplicity, the model demonstrated great success for a number of related sentence-level classification tasks that produced results comparable to those of state-of-the-art. Some of these tasks share similarities with metadiscourse tagging, such as sentiment analysis, in which a tag is assigned to the sentence on the basis of the presence of some words that indicate either *positive*, *negative*, or *neutral* attributes. Hence, for the task of metadiscourse tagging, the use of both CNN and pre-trained word vectors may provide better results than the traditional approach of feature representation with SVMs.

Additionally, the previous chapter has shown other feature types that prove to be effective in detecting and classifying metadiscourse tags, such as POS tags and prosodic features. However, in this context, how to represent these features, particularly POS tags, in continuous space is unclear because POS tag values are considered discrete compared with the continuous values of prosodic cues. One possible solution is to use POS tags with a continuous distribution for each word in the corpus, with corpus-wide information considered (Schmid, 1994). These POS tag distributions, along with other features (*e.g.* pre-trained vectors) in the embedding layer, will then capture the non-linear interactions between the different types of features during the training phase of the network. This approach has demonstrated its effectiveness for the task of POS tagging, as studied by Tsuboi (2014). Tsuboi (2014)'s approach used a hybrid model to take advantage of two types of features, namely discrete and continuous. However, in this work, seeing how a single-layer CNN model can capture non-linear interactions between the pre-trained word vectors and other continuous features, such as POS tags for the metadiscourse tagging task, would be interesting. Aside from this, for the task of metadiscourse tagging, investigating other features, such as prosodic cues in the CNN model, would also be interesting.

6.1.2 MDT-CNN: Overview

This chapter presents an extension of an existing CNN architecture proposed by Kim (2014) and aims to investigate the appropriateness of using such a model for the task of metadiscourse tagging, herein referred to as MDT-CNN. In addition, this chapter shows the non-linear interaction for a combination of features for the task at hand. These features are pre-trained word vectors, POS tag distributions and prosodic cues. The model was evaluated with the same annotated dataset used previously through a 10-fold cross-validation. All the experiments presented in this chapter report only the average results of the 10-fold cross-validation of specific metadiscourse tags. Experiments with all-domain metadiscourse tagging as investigated with MDT-SVMs was neglected in this chapter as its will be often the case that in-domain metadiscourse tagging would provide better results.

Because the settings for the pre-trained word vectors have different variations, these may affect model performance. Hence, a number of experiments were conducted to investigate several variations and thus determine the suitable CNN configurations for the metadiscourse tagging task. These experiments targeted the following two aspects: one related to whether these word vectors are better tuned (non-static) for the task than other model parameters or are kept fixed without tuning (static), and the other is related to the use of different contexts of pre-trained word vectors that were trained to capture local context, as in *word2vec* (Mikolov et al., 2013), to capture global context, such as *GloVe* word vectors (Pennington et al., 2014), or a combination of the two.

The outcomes of these word vector experiments were then used in the settings for the next set of experiments, designed to show the effect of feature combination. In particular, the aim here is to prove the findings from the previous chapter – that the best features for the task are a combination of the above-mentioned features but with the use of CNNs and continuous feature representations instead. All possible combinations of features were tried, to provide an overview of the best feature settings for the task with the use of a CNN model. Finally, these settings were used to configure the network and obtain the final results for the comparison with the SVM model. In addition, an analysis was conducted for the metadiscourse tags to find any areas of improvement for the tagging task in the future.

6.1.3 MDT-CNN: Contributions

The main contributions of the proposed work fall into the following categories:

- Proving the suitability of using CNNs for the task of metadiscourse tagging by using an existing model architecture

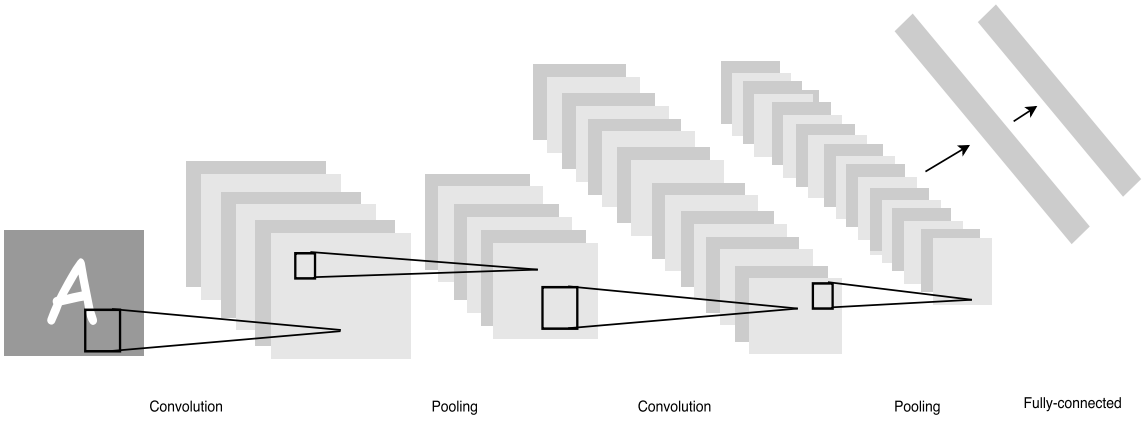


FIGURE 6.1: The CNN architecture of the LeNet-5 model, adopted from (LeCun et al., 2006).

- Studying the effects of using several variations of the pre-trained word embeddings for the task, such as *word2vec* and *GloVe*
- Finding a set of features or feature combinations that are representative of the task in the CNN framework
- Conducting an analysis of the results of the metadiscourse tags to find areas for improvement for the classification model

6.2 Related Work

Initially, CNNs and their components are further explored in Section 6.2.1. This preliminary discussion about CNNs is required to understand how the network architecture is designed from an NLP perspective. Because feature representation is considered an integral part of any NN model, a brief overview of the most prominent methods is given in Section 6.2.2. Finally, a short survey of related sentence-level classification tasks is presented in Section 6.2.3.

6.2.1 Preliminary

Typically, CNNs have a structure similar to that of the LeNet-5 system, which is shown in Figure 6.1. LeNet-5 was developed by LeCun et al. (1998); it is one of the earliest systems based on CNNs in the image processing community, and it is designed for the character

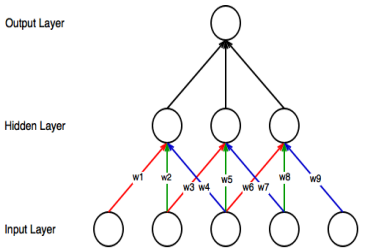


FIGURE 6.2: Three layers CNNs, where each neuron connected to only three adjacent neuron. Edges with same colour share the same weights.

recognition task. Any CNN architecture is designed in such a way that two main properties are introduced, namely local connectivity and shared weights, to guarantee some level of shift, scale and distortion invariance (LeCun et al., 1998). The section below explains each of these in detail with respect to the network design.

- **Local Connectivity**: This is inspired by the organisation of a cat's visual cortex. Early work by Hubel and Wiesel (1968) showed that a cat's visual cortex contains a complex arrangement of cells that are sensitive to small sub-regions, called a receptive field; these sub-regions are tiled to cover the entire visual fields. With the local receptive field, nodes can select the basic features. These features are then combined by the following layer to ensure the identification of high-order features. For instance, each neuron in the hidden layer is forced to have a small number of connections to the adjacent neurons in the input layer (LeCun et al., 1998). This local connectivity is also extended to other layers of the network. This process is illustrated graphically in Figure 6.2, in which each neuron in the hidden layer is only connected to three adjacent neurons in the input layer. The connectivity of neurons in the output layer to the hidden layer is similarly arranged. In this way, the network ensures that it filters only the most informative response from the input (LeCun et al., 1998).

- **Shared Weights**: The concept of weight-sharing is based on the assumption that basic features can be useful across the image despite its position. This can force input units to have similar weights even if their receptive fields are positioned at different locations on the input layer. Hence, the units in the layer are organised in such a way that they all share the same weights. An example would be the edges in Figure 6.2, in which the weights of the same filter are shared across the same layer. Instead of all weights being stored, only w_1, w_2, w_4, need to be stored. With these constraints, the network has an interesting side-effect in that it can be quite compact in terms of the number of actual parameters, and it is perhaps easier to train. All of these weights are learned with the use of the back-propagation process, as explained in Section 6.2.1.

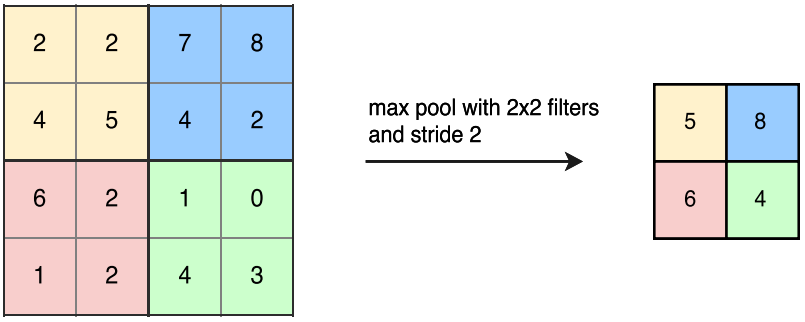


FIGURE 6.3: Demonstrating the max pooling process.

CNNs Components

Typical CNN architecture comprises the following main layers: convolutional layer, pooling layer and fully connected layer. An addition to these layers, the first layer is sometimes called the embedding layer, which receives the network inputs (*e.g.* images of characters or sentences) that are usually normalised in size. In the following, we explain these layers in detail with respect to their order in the network.

Embedding Layer: This is the input layer that contains the embedded vectors, which are low-dimensional representations of each core feature. These vectors can be treated as other network parameters that will be learned during the training phase. This layer is considered a key aspect in any NLP NN architecture. It also contains the embedded vectors as low-dimensional representations of each core feature, as will be discussed in Section 6.2.2.

Convolutional Layer: This is the most important component in CNN architecture and comprises multiple feature maps, each of which has different weight vectors for the selection of different features from all possible locations on the input space (LeCun et al., 1998). For instance, in Figure 6.1, the units in the first hidden layer have six feature maps, and then each unit in the feature map is connected to a 5 x 5 area in the input layer, called the receptive field of that unit. As mentioned previously, all units share the same weight in the feature map for the selection of the same feature in any location on the input. However, other feature maps use different weights to select different local features. A key characteristic of the convolutional layers is that if the input matrix is shifted, then the output of the feature map will be shifted accordingly. By doing so, the network ensures robustness in shifting and dealing with the input. In computing the output of any convolutional layer, its value is first passed to an activation function. Several activation functions can be applied in this case;

the best-known functions are the sigmoid function, hyperbolic tangent (tanh) and rectified linear unit (ReLU) (Nair and Hinton, 2010). Three hyperparameters control the size of the output of this layer:

- **Depth**: This controls the number of neurons in this layer connected to the same region in the input layer.
- **Stride**: This controls the way the width and the height dimensions of the depth are specified. Small values, *e.g.* 1, lead to large output volumes. With a larger value, the output volume will be of smaller dimensions.
- **Zero-padding**: In some cases, padding the input with zeros on the border of the input volume is suitable. This means that any elements in the input that fall outside of the filter region are taken to be zero. Using zero padding is called *wide convolution*; not using zero padding is called *narrow convolution* (Kalchbrenner et al., 2014).

Pooling layer: Sometimes, this is also called *sub-sampling*, which is another key layer in the CNN. The main purpose of this layer is to extract the most important features regardless of their position. The precise locations are not just irrelevant when the features are selected, but they are actually harmful to retain because they signal different locations for the same target, for example the word or character (LeCun et al., 1998). Local averaging, also called average pooling, is used to reduce such precision when the features in the feature maps are selected. For example, in Figure 6.1, the second hidden layer of the LeNET-5 system is performing average pooling. In addition to this, other functions can be applied, such as max-pooling (Huang et al., 2007), which takes the max value of each index instead of the average. An example of the max-pooling process is depicted in Figure 6.3.

Fully Connected layer: This is the final layer in the network after a number of convolutional and max pooling layers. The neurons or nodes in this layer have full connections to all activations in the previous layer, as with regular NNs. In many cases, one is interested in modelling the probability distribution of the output over the possible output classes. This means that the vector of the output needs to be transformed into non-negative real numbers that sum to one, a process that makes the vector a discrete probability distribution over the possible outcomes. This can be achieved through the application of a common transformation function, such as that applied by softmax.

CNNs Training

The objective of training any NN, including CNNs, is to reduce the loss function over the training examples. More precisely, any NN training algorithm shares the same general processes, and these are described as follows:

1. Define a loss function $\mathbf{L}(\hat{\mathbf{y}}, \mathbf{y})$ that quantifies the difference between the predicated $\hat{\mathbf{y}}$ and the true value $\mathbf{y}$.
2. Compute the error frequently over the training examples.
3. Configure the CNN parameters in a way that reduces the loss.
4. Move the parameters in the direction of the gradients.

One of the standard training algorithms is the SGD. A better estimate of the gradients is provided by a large value, whereas a small value has the advantage of converging faster. However, in practice, one can use a graphic processing unit (GPU) to allow efficient parallel computation. The back-propagation algorithm is a method that computes the gradients of the loss function by using the chain rule while keeping records of intermediary results (LeCun et al., 1998, Rumelhart et al., 1988).

6.2.2 Feature Representation

Before the CNN architectures of related classification models are discussed in detail, paying attention to how features (*e.g.* words) are represented in such models is important. In fact, feature representation is one of the key aspects in building any classification model, either linear or non-linear, such as CNNs. The only difference from the linear model is the use of dense vectors instead of sparse input or what is called one-hot representation. The dense representation of words can capture both semantic and syntactic relations between them (Mikolov et al., 2013). Another advantage of using continuous representation is to allow feature generalisation, in which similar words should have similar vectors. This particular advantage of dense representations is important for the modelling task, which involves a similarity between words or phrases, such as the case with metadiscourse tagging. Additionally, dense representations allow feature combination automatically and does not need manual engineering as in the case with sparse representations. Two approaches can be found in the literature on how to derive dense feature representations for the task at hand.

Supervised Pre-training: This method involves the use of a small amount of labelled training data for the intended task (*e.g.* syntactic parsing) and enough labelled data for a related task, called an auxiliary task in this case (*e.g.* part-of-speech tagging). The word vectors are trained first on the auxiliary task, and then the trained vectors are used either as fixed vectors or are further tuned for the intended task.

Unsupervised Pre-training: This is the most common method, in which no auxiliary task with enough labelled data for training is involved. In this method, the word embedding vectors are trained with the use of a large amount of unlabelled data for auxiliary tasks, such

as the language model. Some of the widely used techniques to perform the unsupervised approach are *word2vec* (Mikolov et al., 2013), *GloVe* (Pennington et al., 2014) and Collobert and Weston's technique (Collobert and Weston, 2008, Collobert et al., 2011). These models that generate a continuous feature representation of words are also referred to as *neural language* or *word embedding* models, and they are trained with the use of an SGD algorithm. Other discrete features, such as POS tags and named entity tags, can be concatenated to the word vectors and trained to produce dense vectors that are representative of these features.

The choice of context poses an important factor in these models, in which the window can be set up to be around the word in question or within the same sentence, paragraph or document. The most common approach is a sliding window, in which language models are built by looking at a sequence of $2j+1$ words. The middle word is called the focus word, and the j words to each side are the contexts. Also, the size of the sliding window strongly affects the resulting vectors, in which larger windows tend to produce topically more similar vectors, whereas smaller window sizes tend to produce more functional and syntactic similarities.

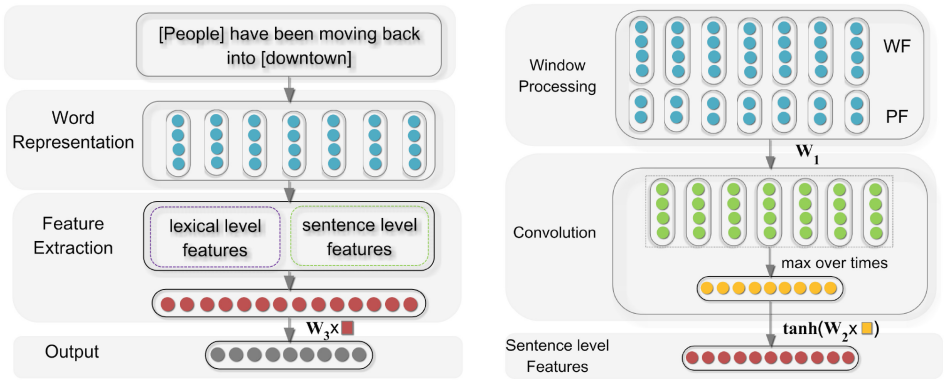
6.2.3 CNNs Architectures

The previous sections have provided the required background to understand CNNs in general. However, the related structure varies for different tasks because one can have single or multiple convolutional layers. In the following section, studies in the literature on the use of CNNs as classification models are grouped on the basis of the number of convolutional layers, namely single layer or multiple layers.

Single Convolutional Layer

For the task of relation classification, Zeng et al. (2014a) utilised a single-layer CNN to extract lexical and sentence-level features. Initially, word vectors are prepared with the use of word embeddings, and then the lexical features are extracted with respect to the given nouns. At the same time, sentence-level features are learnt during the training phase of the network. Finally, these two types of features are concatenated together and fed to the softmax classifier for predictions. The experiment was conducted with the use of SemEval–2010 Task 8 dataset (Hendrickx et al., 2009), and the reported results outperform those of state-of-the- art methods.

Shen et al. (2014) and Yih et al. (2014) both proposed similar CNN architectures, which first convert each of the word tokens into vectors with the use of a letter-tri-gram in the embedding layer and then use these vectors as inputs to the convolutional layer to extract local features. Finally, the max over-time pooling layer is used to form a global feature



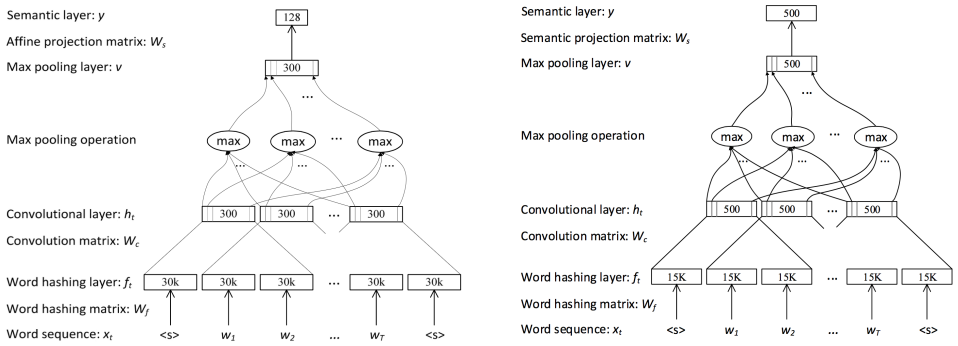
(A) (B)

FIGURE 6.4: Architecture of the neural network used for relation classification illustrated in (A), The framework used for extracting sentence-level features presented in (B) Zeng et al. (2014a).

vector; a fully connected layer serves as the output layer, and a final semantic layer is used to represent the high-level semantic feature vector of the input word sequence, as shown in Figure 6.5. However, the model (Figure 6.5 (A)) in the work of Shen et al. (2014) was proposed for the task of web searches, in which the aim was to test the model on a question set from a commercial search engine. By contrast, the model (Figure 6.5 (B)) in the work of Yih et al. (2014) used an open-domain QA dataset to train one model for relation extraction and another for entity extraction. The model was defined as a multi-class classification, in which, for example, the top 150 candidates are returned when a query is given.

Kim (2014) proposed a CNN architecture for several sentence-level classification tasks. The model is composed of one convolutional layer, followed by a max over-time pooling, and a fully connected layer with dropout and softmax output layers. The convolutional layer consists of three parallel convolutional layers with different filter sizes, as shown in Figure 6.8. The model has two channels, namely one that takes the input of randomly initialised word vectors and the other that uses *word2vec* embedding vectors. However, only the parameters of one channel are updated during the training phase of the network, and this case is called 'non-static'. The static case means keeping the model parameters fixed during the network training. The model has the ability to carry out both binary and multiclass classification, and reports excellent results for a series of sentence classification tasks, and using different datasets, as follows:

- MR: Movie reviews have one sentence per view as positive or negative (Pang and Lee, 2005).



(A) (B)

FIGURE 6.5: The Similar CNNs architectures of Shen et al. (2014) in (A) and Yih et al. (2014) in (B)

- SST-1: The Stanford Sentiment Treebank is considered an extension of MR but with fine-grained labels – very positive, positive, neutral, negative, very negative (Socher et al., 2013).
- SST-2: This is the same dataset as SST-1 but uses positive/negative labels only.
- Subj: This is a subjectivity dataset, in which the task is to classify a sentence according to subjective or objective labels (Pang and Lee, 2004).
- TREC: The Text REtrieval Conference question dataset has six labels – whether the question is about a person, location, numeric information, *etc.* (Li and Roth, 2002)
- CR: These are customer reviews with positive/negative labels (Hu and Liu, 2004).
- MPQA: This is an opinion dataset with positive/negative labels (Wiebe et al., 2005).

Multiple Convolutional Layer

A CNN model was proposed by Hu et al. (2014) for sentence matching tasks. The structure of the model consists of a 1D convolution, followed by a 1D max-pooling, several 2D convolution and pooling layers as well as several fully connected layers, as depicted in Figure 6.6. The inputs to the network are word embeddings, which are trained with *word2vec* (Mikolov et al., 2013); other embeddings were learnt with the use of Wikipedia (1B words) for English words and Weibo data (300M words) for Chinese words. The model does not require any prior knowledge of the language and claims applicability to any other matching tasks. The model has been validated in several tasks and hence datasets, such as sentence completion (Lewis et al., 2004), matching a response to Weibo, and the MSRP dataset Vasile et al. (2008). The experimental results gave superior results for the task of sentence matching.

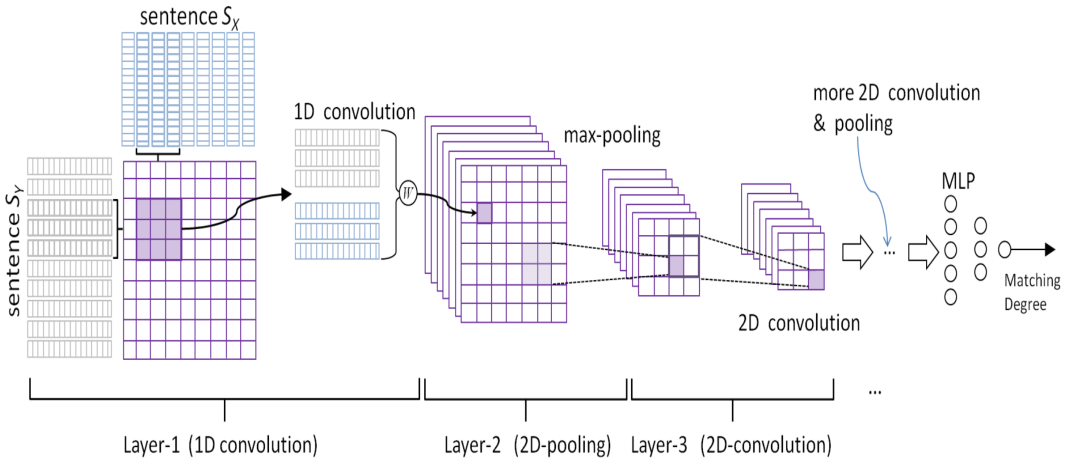


FIGURE 6.6: Hu et al. (2014) model architecture with two convolutional layers for sentence matching task between sentence S_x and S_y.

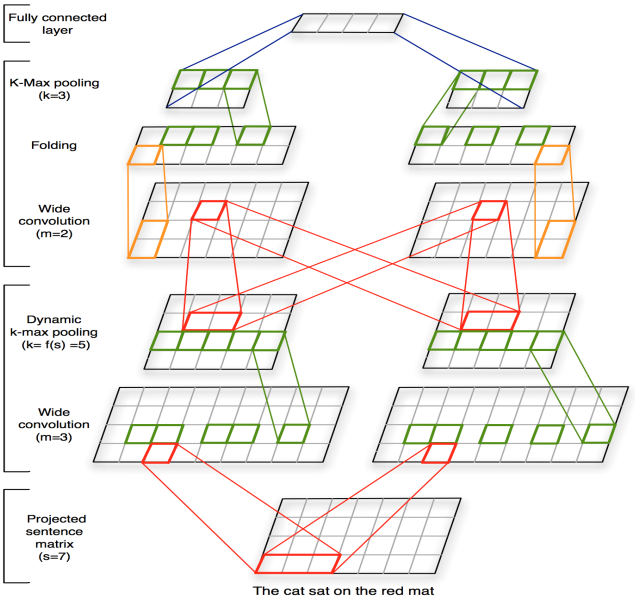


FIGURE 6.7: Kalchbrenner et al. (2014) model architecture using k-max pooling.

Kalchbrenner et al. (2014) presented a dynamic k-max pooling CNN (DCNN) for sentence modelling. The model structure is composed of a number of wide 1D convolution layers, followed by a feature map folding operation and k-max pooling layer, as well as a fully connected layer as output, which is illustrated in Figure 6.7. The concept of k-max pooling operation enables the pooling of the k most active indicators that may be a number of positions apart; it preserves the order of the features, but it is insensitive to their locations. The model has been validated on several datasets, such as SST-1, SST-2, six-type question categorisation in the TREC dataset and Twitter sentiment prediction tasks (tweets with positive/negative labels). Dynamic k-max pooling CNNs outperform the single-layer CNNs

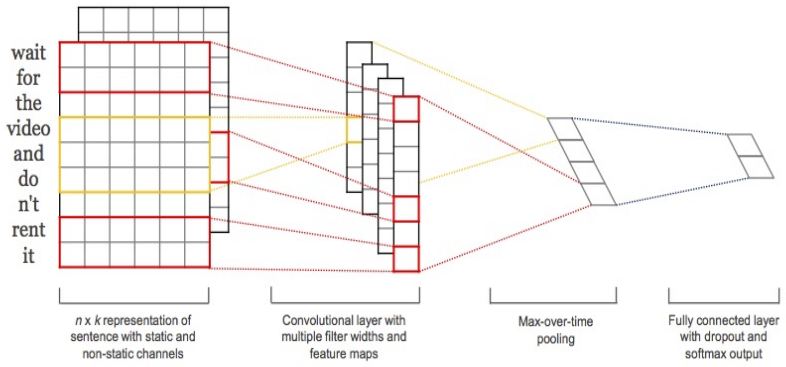


FIGURE 6.8: Kim (2014) model architecture with two channels for an example sentence.

proposed by Kim (2014) on SST-1 and TREC, but not on SST-2, in which the model of Kim (2014) performed better.

Despite the simple structure of the CNNs model proposed by Kim (2014) in different applications in several classification tasks, they produce state-of-the-art results. In this work, the multi-class classification model developed by Kim (2014) is adopted for the task of metadiscourse tagging. The aim here is two-fold: first, to investigate the suitability of the CNN for the metadiscourse tagging, and second is to show the effects of continuous feature representations on the classification task.

6.3 CNNs-based Metadiscourse Tagging

This section presents the metadiscourse tagging using CNN (MDT-CNN) model that considers both text-based and acoustic-based features in the continuous space. The MDT-CNN model consists of three parts: continuous feature representation for words, POS tags and prosodic cues (PRO); a multi-class classification using a single-layer CNN; and, finally, a regularisation process to optimise the networks during training.

6.3.1 Features

The previous chapter has shown the effectiveness of using a combination of n-grams, POS tags and PRO. One of the objectives for using a CNN model is to validate these findings. Thus, the same set of features is explored in this chapter, but here they are presented in the continuous space. As the set of PRO are real values, they will be used as was explained in Chapter 4. This set includes F0 and pause duration, which means that there are two

dimensions in the feature space. Hence, this section is only concerned with the other features, namely pre-trained word embedding and POS tag distributions.

Word Embeddings

The well-known pre-trained embedding vectors *word2vec* (Mikolov et al., 2013) and *GloVe* (Pennington et al., 2014) were selected as sources of word continuous features. These vectors were trained using NN models, which are often referred to as neural language models (Bengio et al., 2003). The window contexts in such models are either global or local.

word2vec relies on local window contexts and has two forms: skip-grams or continuous bags-of-words (CBOWs). In skip-grams, if the window size is c around the target token t, its predication for the contextual words is $p(c|t)$, while CBOWs make predictions for the current token t, given its context as $p(t|c)$. The pre-trained *word2vec* used is based on CBOWs, trained on 100 billion words from Google (Mikolov et al., 2013) and has a dimensionality of 300. *GloVe* is based on a global log-bilinear regression model that combines the advantages of both local and global contexts and is trained on 840 billion tokens with 300 dimensions (Pennington et al., 2014).

In this work, the focus is on the local context, as this is more appropriate for the tagging task at the sentence level. In addition, the use of *word2vec* has produced successful results in related tasks that need local contexts, as well as in sentiment analysis (Kim, 2014) and dialogue acts tagging (Kim et al., 2015, Milajevs et al., 2014). However, the global context could also be useful for this task since it captures the structure of an entire document, which might have an impact on the model's performance. Thus, both sources *word2vec* and *GloVe* were considered for the proposed task.

Despite the effectiveness of using such pre-trained vectors in related work, there is one issue that often arises, which is the out-of-vocabulary (OOV) problem. That is, there may be words in the dataset that have no representations in the pre-trained word embeddings. To address this issue, the vectors of the OOV words were randomly initialised.

POS Tags Distributions

It is important to clarify how to convert POS tags from discrete features to continuous space. For this purpose, the method proposed by Schmid (1994) and Tsuboi (2014) was followed to produce POS tag distributions over the training data. That is, each word token t was represented by a vector size of $|Q|$, where the q^{th} element was the conditional probability with which that word gets the q^{th} POS tag. Thus, the probabilities for a POS tag q were estimated with additive smoothing as:

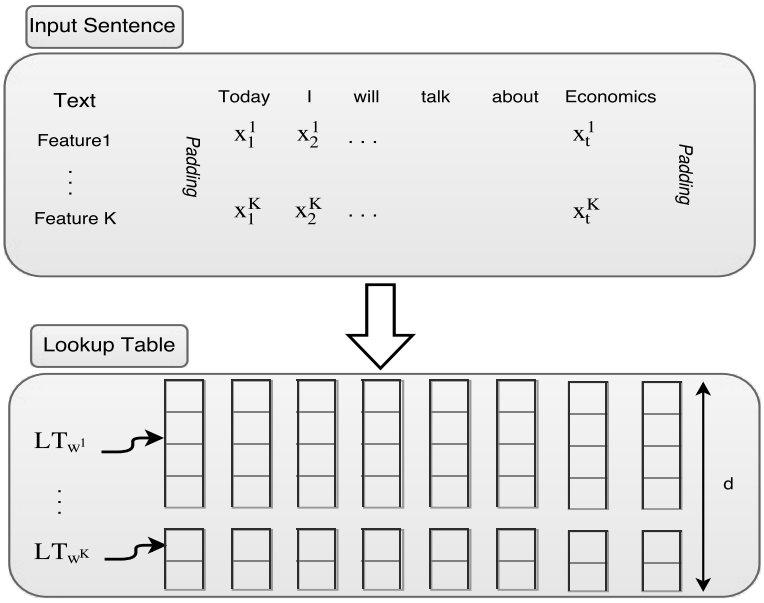


FIGURE 6.9: The structure of the Lookup Table adopted from (Collobert et al., 2011). Padding refer to process of having fixed size of sentences across the whole corpus.

$$P(q|t) = \frac{C(t,q)+1}{C(t)+|Q|}, \tag{6.1}$$

where $C(t)$ and $C(t,q)$ were the counts of t and the co-occurrences of t and q, respectively. The $|Q|$ was then concatenated with the other features as explained below.

With regards to the representation of the POS tag distributions and PRO, these were appended to the pre-trained word vectors in a manner similar to the one presented in Figure 6.9, as examined by Collobert et al. (2011). Then, as with the word vectors, those features were tuned during the training phase of the network.

More formally, consider word token t and that each token is represented by K features $[t \in \mathcal{D}^1 \times ... \times \mathcal{D}^K]$, where $\mathcal{D}^k$ is the dictionary for the k^{th} feature. A lookup table, $LT_{x^k}(.)$ is associated with each feature, with parameters $x^k \in \mathbb{R}^{d^k_{emb} \times |\mathcal{D}^k|}$, where $d^k_{emb} \in \mathbb{S}$ is the total vector size. In this case, this is equal to 338 (300 for word embeddings, 36 for POS and 2 for PRO cues) in our case. Given a token t, a feature vector of dimension $d_{emb} = \sum_k d^k_{emb}$ is then obtained by concatenating all lookup table outputs:

$$LT_{x^1,...,x^K}(t) = \begin{pmatrix} LT_{x^1}(t) \\ \vdots \\ LT_{x^K}(t) \end{pmatrix} = \begin{pmatrix} x^1_t \\ \vdots \\ x^K_t \end{pmatrix}.$$

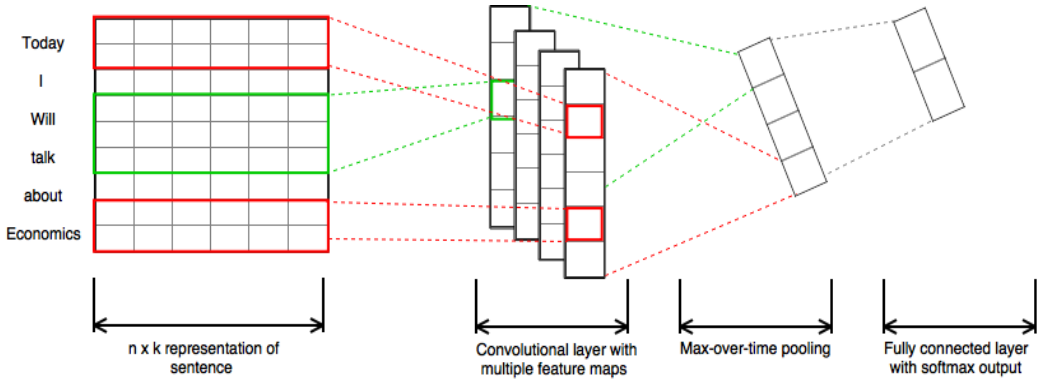


FIGURE 6.10: The used CNNs Architecture which was adopted from (Kim, 2014) with only one channel.

The final matrix output that contains all the lookup features tables for all the tokens in the sequence $t_1, ..., t_j$ is as follows:

$$LT_{x^1,...,x^K(t_1,...,t_J)} = \begin{pmatrix} x^1_{t_1} & ... & x^1_{t_J} \\ & \vdots & \\ x^K_{t_1} & ... & x^K_{t_J} \end{pmatrix}.$$

For simplicity, the previous matrix could look like

$$LT_{\mathbf{X}}(t_1, ..., t_j) = (\mathbf{x}_1 \ ... \ \mathbf{x}_j). \tag{6.2}$$

Where $\mathbf{X} \in \mathbb{R}^{d_{emb}}$ is a matrix of parameters that need to be tuned during the training process, $\mathbf{x}_j$ is the j^{th} column of $\mathbf{X}$ that corresponds to the feature vector of the token t_j. d_{emb} is the token vector size (338) Then, this matrix is fed into other layers of CNNs, as we will see below.

6.3.2 Model Architecture

This section describes the extraction of the most informative features for classifying metadiscourse tags using the previously prepared input matrix X from the embedding layer. It also shows how the components of the CNN architecture fit together to serve this aim. As stated earlier, the model architecture is a one-layer CNN adopted from the work of Kim (2014), but instead of using two channels for the input, only one was used, as illustrated in Figure 6.10. Despite its simplicity, the model achieved state-of-the-art results when used as a replacement for existing text classification baselines, such as SVM and logistic regression, for a number of sentence-level classification tasks.

To further explain the CNN model used, consider a sentence of j tokens $t_1, ..., t_j$ (zero-padding where necessary), each with their corresponding d_{emb} dimensional embedding vector $\mathbf{x}_j \in \mathbb{R}^{d_{emb}}$. These vectors are the outputs of the embedding layer that was described in Equation 6.2; for such sequences, the representations are:

$$\mathbf{x}_{1;j} = \mathbf{x}_1 \oplus ... \oplus \mathbf{x}_j, \tag{6.3}$$

where $\oplus$ denotes the concatenation operator between feature vectors. Then, this sentence matrix can be treated as an image of size $j \times d_{emb}$ (Collobert et al., 2011). This matrix is further fed into a convolution layer, which involves a linear filter, $\mathbf{w} \in \mathbb{R}^{hd_{emb}}$, that is applied over the sentence by moving a sliding window of size h words that corresponds to the height of the filter (simply called the 'region size"; Zhang and Wallace (2015)) to extract the most salient information in this region. Thus, the filter is parameterised by the weight matrix $\mathbf{w}$ with region size h, and w contains $h.d_{emb}$ parameters to be estimated. For example, feature u_i is generated from a window of words $\mathbf{x}_{w_{i:i+h-1}}$ by

$$u_i = g(\mathbf{w} \,.\, \mathbf{x}_{w_{i:i+h-1}} + b). \tag{6.4}$$

Here, b is a bias term and g is a non-linear function (*e.g.*, the hyperbolic tangent). This filter is applied to each window over the sentence $\{\mathbf{x}_{w_{1:h}}, \mathbf{x}_{w_{2:h+1}}, ..., \mathbf{x}_{w_{j-h+1:j}}\}$ to produce a *feature map*

$$\mathbf{u} = [u_1, u_2, ..., u_{j-h+1}]. \tag{6.5}$$

In theory, when multiple feature maps are generated, each one has different dimensionality as the sentences length is varied. To fix this issue, these feature maps are fed into the pooling layer, where a max-over-time pooling operation (Collobert et al., 2011) is applied by taking the maximum value $\hat{u} = \max\{\mathbf{u}\}$. Ideally, the process would be run more than once to generate multiple filters. The outputs from each filter are then concatenated to form a fixed-length feature vector that is more representative for the sentence. That is, by applying m filters, the generated features vector is defined as $\mathbf{z} = [\hat{u}_1, ..., \hat{u}_m]$. Finally, the vector $\mathbf{z}$ is passed to the fully-connected layer to be used for the final classification of multiple metadiscourse tags using a softmax function.

6.3.3 Regularisation

As the training objective was to minimise loss L, all the network parameters (including the weight of the filter, the bias term in the activation function and the embeddings E) were set to this objective. However, these parameters are prone to overfitting. One way to solve this issue is by using the dropout method, which is a regularisation method that is often used with a constraint on l_2-norms of the weight vectors to prevent co-adaptation of hidden units (Hinton et al., 2012). The method works by randomly dropping out (*i.e.* setting to zero) a proportion p of the hidden units during back-propagation (Kim, 2014). As demonstrated by Kim (2014), the network output is computed in the standard forward propagation:

$$\bar{\mathbf{y}} = \mathbf{w} \,.\, \mathbf{z} + b \tag{6.6}$$

However, by imposing the dropout, it becomes

$$\bar{\mathbf{y}} = \mathbf{w} \,.\, (\mathbf{z} \circ \mathbf{r}) + b, \tag{6.7}$$

where $\circ$ is the element-wise multiplication operator and $\mathbf{r} \in \mathbb{R}^m$ is a 'masking' vector of Bernoulli random variables with a probability p (Kim, 2014). This way, the back-propagated gradients are passed only through the unmasked units. During the test phase, the weight vectors are scaled by an amount of p in which $\hat{\mathbf{w}} = p\mathbf{w}$. Then, $\hat{\mathbf{w}}$ is used (no dropout is used in this stage) to score unseen examples (i.e. sentences) (Kim, 2014). In addition, $l_2 - norms$ of the weight vectors are constrained after a gradient descent step, by rescaling $\mathbf{w}$ in order to have $||\mathbf{w}||_2 = s$, when $||\mathbf{w}|_2| > s$.

Finally, to compute the loss one needs to measure the dissimilarity between the true label distribution $\mathbf{y} = y_1, ..., y_j$ and the computed network's output $\bar{\mathbf{y}}$. However, $\bar{\mathbf{y}}$ need to be transformed to $\hat{\mathbf{y}} = \hat{y}_1, ..., \hat{y}_j$ using the softmax activation function to produce a non-negative discrete probability distribution over C classes that sum to 1; that is, by using the cross-entropy loss

$$L_{cross-entropy}(\hat{\mathbf{y}}, \mathbf{y}) \;=\; -\sum_i \frac{1}{n_i}\, y_i \, log(\hat{y}_i), \tag{6.8}$$

where n_i is the total number of samples in class i. This weight is used to control the imbalanced distributions between classes.

Description	Values
input word vectors	Google word2vec
filter region size	(3,4,5)
feature maps	100
activation function	ReLU
pooling	1-max pooling
dropout rate	0.5
l2 norm constraint	3

TABLE 6.1: CNNs Baseline configuration. 'feature maps' refers to the number of feature maps for each filter region size.

6.4 Experiments and Results

6.4.1 Experimental Setup

CNNs Settings

It is worth noting that the MDT-CNN model was trained using SGD over shuffled mini-batches, which were set to 50, using the *AdaDelta* (Zeiler, 2012) update rule. While the SGD algorithm can, and often does, produce good results, AdaDelta is designed to select the learning rate for each minibatch, sometimes on a per-coordinate basis. Moreover, it has been proved that as a network structure becomes more complex, computing speed can be affected by limited computing resources (Steinkrau et al., 2005). Thus, a graphics processing unit (GPU) was used to enable a more efficient computational process (Steinkrau et al., 2005). Table 6.1 shows the hyperparameter configuration of the MDT-CNN model, which is comprised of the same settings reported by Kim (2014). Experiments were performed with the Theano python toolkit (Bastien et al., 2012, Bergstra et al., 2010) using a NVIDIA K20 GPU.

Evaluation Procedure

As explained in Section 6.3, the MDT-CNN model was developed as one classifier for all 20 specific metadiscourse tags, including the NONE tag which refers to cases where the utterances have no instances of metadiscourse. In addition, the evaluation settings (*e.g.* dataset and evaluation metrics) that were used for MDT-SVM and explained in Section 5.4.1, were also used here for the MDT-CNN model. For instance, the model was trained and tested using a 10-fold cross-validation process on the same gold-standard dataset described in Table 5.2. The evaluation metrics used were precision, recall and F.

Training Mode	Physics			Economics			Average		
	P	R	F	P	R	F	P	R	F
Static	60.73	32.85	42.64	59.19	40.30	47.95	59.96	36.58	45.29
Non-static	60.00	41.35	**48.96**	62.77	45.68	**52.88**	61.39	43.52	50.92

TABLE 6.2: Results of MDT-CNN model showing the effects of static and non-static word embeddings using *word2vec* on the model performance on the two disciplines. Bold face denotes statistically significant results.

Word Vectors	Physics			Economics			Average		
	P	R	F	P	R	F	P	R	F
word2vec	60.00	41.35	**48.96**	62.77	45.68	**52.88**	61.39	43.52	50.92
Glove	57.30	38.55	46.09	61.57	44.88	51.92∗	59.44	41.72	49.01
word2vec+Glove	54.55	37.65	44.55	59.47	42.98	49.90∗	57.01	40.32	47.23

TABLE 6.3: Results of MDT-CNN model using different pre-trained word embeddings strategies. Bold face denotes statistically significant results and * denotes insignificant difference.

6.4.2 Word Embeddings

Word embedding is considered the main features vector used in developing a MDT-CNN model for a tagging task. Therefore, the experiments in this section were organised with respect to two strategies. The first set of experiments investigated the appropriate word embedding settings for the metadiscourse tagging task. Static settings kept the word vectors fixed, including the OOV ones, during the network training, whereas non-static settings tuned the word vectors for each task. These experiments only used *word2vec* to show the effects. The other experiments showed which pre-trained word embeddings work best for the task.

Table 6.2 presents the results of the two settings, using *word2vec* for both disciplines. The results suggest that fine-tuning the word vectors during training significantly improves the MDT-CNN model performance. For instance, when the words vectors were set as non-static, the F score obtained was 50.92, compared to 45.29 with the static case. Similar observations were noted by Kim (2014) on various text datasets for the task of sentence classification. However, that study reported that in some cases the static approach produced better results. This indicates that the decision of whether to use the static or non-static word embeddings settings is task-dependent.

The other set of experiments with regard to the use of different sources of pre-trained word vectors is presented in Table 6.3. The best results were obtained when *word2vec* was used, as on average the F score was 50.92, compared to 49.01 when *GloVe* was used instead. This indicates that word vectors that were trained with respect to the local context (i.e. *word2vec*) outperformed those trained with respect to both the global and local contexts of the documents (*GloVe*). This finding can be attributed to the nature of the metadiscourse

Feature	Physics			Economics			Average		
	P	R	F	P	R	F	P	R	F
LEX	60.00	41.35	48.96	62.77	45.68	52.88	61.39	43.52	50.92
LEX+POS	60.20	42.43	49.78	67.90	47.38	55.81	64.05	44.91	52.79
LEX+PRO	60.65	41.95	49.60	63.27	46.18	53.39	61.96	44.07	51.49
LEX+POS+PRO	60.45	42.93	**50.21**	70.58	48.79	**57.70**	65.52	45.86	53.95

TABLE 6.4: Results of MDT-CNN model in reference transcripts showing the difference in performance when different features are added to the model. All the features are used in non-static mode. LEX denotes the *word2vec* word vectors, *POS* refers to the Part-of-Speech tags distributions and PRO denotes the prosodic cues. Bold face denotes statistically significant results.

tagging task, as it aims to extract the features most representative of a tag at the sentence level. Therefore, using word vectors that take the global context into consideration is less useful in this context. However, it should be noted that the results of *Glove* were not that bad, as the difference between them was statistically insignificant for economics lectures, though not for physics lectures. This may be related to the fact that the model used to produce these vectors took advantage of both global and local contexts, as explained in Section 6.4. Furthermore, simply concatenating the two sources of word embedding, resulting in a vector dimension of 600 (300 for each word vector), had no effect on the results. This was also noticed by Zhang and Wallace (2015), who compared the performance of several datasets for a sentiment analysis task.

The outcomes from this set of experiment reveals two important settings that are suitable for metadiscourse tagging when pre-trained vectors are used. These settings are the importance of tuning the word vectors during the training process and the effectiveness of using *word2vec* as a source of pre-trained vectors. These outcomes will be used for all other experiments presented in this chapter.

6.4.3 Features Combinations

This section reports on a number of experiments that were performed with respect to other features, such as POS tag distribution and PRO, whose effectiveness was shown with the MDT-SVM model in Chapter 5. This step was important to validate the usefulness of this feature combination, as was the case with the MDT-SVM. Note that, based on the outcomes of the previous section, all the features vectors used were tuned when conducting the set of experiments here (*i.e.* a non-static setting). Moreover, these experiments that investigated different features were performed by keeping *word2vec* as the main feature and then either varying the inclusion of the other features (POS tag distribution or PRO) or combining them all.

Feature	Physics			Economics			Average		
	P	R	F	P	R	F	P	R	F
LEX	51.47	34.53	41.33	56.74	40.26	**47.10**	54.11	37.39	44.22
LEX+POS	48.69	32.81	39.20	56.82	36.88	44.73∗	52.76	34.85	41.97
LEX+PRO	50.05	36.80	42.41	59.68	38.19	46.58∗	54.87	37.49	44.49
LEX+POS+PRO	51.03	37.52	**43.24**	59.01	37.64	45.96∗	55.01	37.58	44.60

TABLE 6.5: Results of MDT-CNN model on ASR outouts showing the difference in performance when different features are added to the model. All the features are used in non-static mode. LEX denotes the *word2vec* word vectors, *POS* refers to the Part-of-Speech tags distributions and PRO denotes the prosodic cues. Bold face denotes statistically significant results and * denotes insignificant difference between the results.

Table 6.4 presents the results for all possible combinations of features using the reference transcripts. The first row shows the results of using the *word2vec* which has been obtained from the previous section, herein referred to as LEX and considered the baseline feature in all other experiments of the MDT-CNN. The second row shows the performance of the MDT- CNN model when both LEX and POS tag distributions are used. For both disciplines, such a combination clearly improves the model, with roughly 2% absolute in the $F1$-score average score compared to the case when only LEX features were used. The third row shows the results of the model when using both LEX and PRO. Adding PRO achieved an average $F1$-score of 51.49, which is relatively weak compared to the score obtained when POS tags were used. This finding was consistent in both disciplines; however, its effect negatively impacted the performance in the economics lectures compared to physics, with the $F1$-score dropping from 55.81 (with POS) to 53.39 (with PRO). This can be attributed to the fact that acoustic-based features in general, including PRO, are more variant and speaker-dependent. This observation was also noticed with the MDT-SVM model. However, combining all features together (*i.e.* LEX, POS and PRO) improved the model performance significantly compared to the other features, and this was consistent in both disciplines.

On the other hand, Table 6.5 shows the results for the same set of experiments for feature combinations when ASR transcripts were used. In general, there was substantial degradation in the model performance across both categories and disciplines compared to when the reference transcription was used. Adding POS tag distribution features decreased the model performance in both disciplines, with an average F score of 44.22 to 41.97. This finding is intuitively correct, as the grammatical structure of the sentence plays a vital role in classifying the metadiscourse tags, and losing such a structure, as in the case with ASR output, would have a significant impact on model performance. This observation was also seen in the MDT- SVM model in Chapter 5 when testing the model on ASR outputs. However, adding prosodic features seemed to slightly improve the model performance. Additionally, combining all the features together did not have a substantial effect on boosting the model

Model	Physics			Economics			Average		
	P	R	F	P	R	F	P	R	F
	REF								
SVMs	46.09	42.25	44.09	50.31	47.36	48.79	48.20	44.81	46.44
CNNs	60.45	42.93	**50.21**	70.58	48.79	**57.70**	65.52	45.86	53.95
	ASR								
SVMs	34.49	28.76	31.37	35.66	32.34	33.92	35.08	30.55	32.65
CNNs	50.05	36.80	**42.41**	59.68	38.19	**46.58**	54.87	37.49	44.49

TABLE 6.6: Comparison between CNNs model and SVMs. Bold face denotes statistically significant results.

performance further, as the difference in terms of the F score between this combination and when only LEX was used was insignificant.

6.4.4 Compare CNNs to SVMs

As the metadiscourse tagging task is relatively new, there are unfortunately no direct comparisons with an equivalent state-of-the-art method. Thus, in this section, the results for the best configuration of the MDT-CNN model (*e.g.* non-static feature vector combinations of words, POS tags distributions and PRO) were compared with those using an MDT-SVM model from Chapter 5. For both models, the experiments were evaluated in exactly the same fashion (*i.e.* the dataset used and the evaluation procedures using 19 specific metadiscourse tags for reference transcriptions). Any reporting of results in this section is based on the F score.

Table 6.6 shows the difference in performance between these two models, namely MDT-CNN and MDT-SVM. The first part presents the results on the reference transcriptions, which show that there was a 7.51% significant improvement over the baseline (MDT-SVM) in the $F1$-score. This confirms the hypothesis that the joint modelling of continuous features and CNNs is more suitable for the task of metadiscourse tagging than the traditional approach (*i.e.* MDT-SVM). This remarkable improvement was also noticed in similar tasks, such as sentiment classification as studied by Kim (2014), where such a task shares a common principle with the metadiscourse tagging task. For example, depending on the content of the sentence, it is assigned one of these categories: *positive*, *negative* or *neutral*. As stated before, the model used in this work was actually based on the work of Kim (2014); however, their study only considered pre-trained *word2vec* vectors as their feature set. In the current work, the network was configured to deal with other feature sets, such as POS tag distribution and PRO, which demonstrated effectiveness in boosting the model performance for the tagging task on academic lectures.

	Physics			Economics			Average		
Tag	P	R	F	P	R	F	P	R	F
ML	70.61	62.84	66.50	75.58	47.06	58.00	73.09	54.95	62.25
DO	68.95	56.59	62.16	78.93	65.16	71.38	73.94	60.88	66.77
SA	77.44	64.46	70.35	78.58	71.82	75.04	78.01	68.14	72.69
IA	66.85	37.66	48.18	90.82	58.94	71.49	78.84	48.30	59.84

TABLE 6.7: Results for generic tags metadiscourse Tagging.

However, it was also noticed from the results that both models had similar behaviour on per-discipline performance, as both models performed better in economics than in physics. For example, the $F1$-scores in the physics lectures were 44.09 and 50.21 for the MDT-SVM and MDT-CNN models, respectively. This is in contrast with economics lectures, where the F1-scores were 48.79 for MDT-SVM and 57.70 for MDT-CNN. Such observations were also noticed in the case when ASR outputs are used in both models, as shown in the bottom portion of Table 6.6. This may be due to the fact that the economics lecturers in our dataset more often used slides to guide the students through the lecture content. Such content organisation might have also have had an indirect impact on the use of metadiscourse expressions within the lectures. Further analysis on generic metadiscourse tags confirmed the same observation, as shown in Table 6.7. Most of these generic tags performed better in economics than in physics. However, this was not the case for the *Metalinguistics* (ML) tags, as the F score in physics of 66.50 was better than in economics (58.00). This suggest that such findings are not conclusive and that a more in-depth analysis of the performance of the lower level of metadiscourse tags is needed; *i.e.* specific metadiscourse tags. This is because there are some factors that may affect model performance, such as the number of occurrences of the tags in the datasets or the way the annotation has been done.

6.4.5 Analysis and Discussion

Table 6.8 shows the results of the 19 metadiscourse tags in the gold-standard datasets using MDT-CNN. In the following section, the analysis of these specific metadiscourse tags is organised based on the generic tags that these specific tags belong to. As the case with MDT-SVM, the main observation here is that the frequency of a metadiscourse tag has an impact on model performance. Hence, the following discussion is devoted only to metadiscourse cases where this normal expectation did not seem to apply.

		Physics			Economics			Average		
	Tag	P	R	F	P	R	F	P	R	F
ML	REP	85.92	74.39	79.74	90.36	82.42	86.21	88.14	78.41	82.95
	REF	85.56	96.39	90.65	88.46	82.14	85.19	87.01	89.27	87.92
	CLF	00.00	00.00	00.00	50.00	02.94	05.56	25.00	1.47	2.78
	CLA	50.78	37.40	43.08	77.59	65.22	70.87	64.19	51.31	56.98
	MAT	51.84	42.61	46.78	50.00	21.68	30.25	50.92	32.15	38.52
DO	INT	61.25	26.34	36.84	72.29	57.88	64.29	66.77	42.11	50.57
	CON	55.56	11.63	19.23	75.76	26.04	38.76	65.66	18.84	28.99
	DEL	88.89	11.59	20.51	70.83	25.76	37.78	79.86	18.68	29.15
	COT	00.00	00.00	00.00	100.00	04.76	09.09	50.00	2.38	4.55
	ENU	54.46	40.52	46.47	66.67	46.90	55.06	60.57	43.71	50.77
	PHO	59.74	39.32	47.42	71.07	61.08	65.70	65.41	50.20	56.56
	REV	74.73	64.01	68.96	76.13	68.30	72.00	75.43	66.16	70.48
	PRE	56.30	38.43	45.68	72.51	61.38	66.48	64.41	49.91	56.08
SA	EMP	67.43	52.16	58.82	68.49	60.05	63.99	67.96	56.11	61.41
	EXE	84.29	80.88	82.55	86.25	87.85	87.04	85.27	84.37	84.79
	ARG	97.06	84.62	90.41	00.00	00.00	00.00	48.53	42.31	45.21
	SUG	00.00	00.00	00.00	88.89	38.10	53.33	44.45	19.05	26.67
IA	MAC	74.04	36.67	49.04	91.84	83.33	87.38	82.94	60.00	68.21
	AAR	68.42	24.53	36.11	20.00	02.33	04.17	44.21	13.43	20.14

TABLE 6.8: Results of CNNs model for specific tags Tagging.

Metalinguistics Comments (ML)

Both *Repairing* (REP) and *Reformulating* (REF) tags have unusual performance, as the average F1-scores were 82.95% and 87.92%, respectively. However, they had fewer occurrences in both disciplines (only 173 in REP and 244 in REF), as indicated in Table 5.2 in Section 5.4.1. This is in contrast to the *Managing Terminology* (MAT) tag, where the performance was much lower (38.52%) despite higher occurrences (817) in both disciplines. This suggests that there is a need for further inspection of such behaviour, as this might be related to the way the annotators marked these tags. Another possible reason is that these particular tags may exhibit more variation and the model may need to include more examples of such expressions to be able to accurately identify and classify them.

Discourse Organisation (DO)

The PHO category among the set of tags of DO occurred only 298 times, but its performance was comparable to the performance of PRE, which had more occurrences in the dataset

(802). This observation was noticed in both disciplines, as the F1-scores for PHO and PRE in Physics were 47.42 and 45.68, respectively, and 65.70 and 66.48, respectively, for Economics. This can be attributed to the fact that lexical expressions that are used to indicate PHO, such as "This is the slide" or "Look to that figure", are much clearer and somehow semi-fixed compared to those used for PRE, which is made up of hypotheses that exhibit a lot of variation, as also noticed with MDT-SVM in Chapter 5. Again, far more examples may be needed to be able to identify and classify the PRE tag. Another possible reason that was revealed when manually inspecting the instances of PRE in the gold dataset is that the expressions used to indicate this category can be split into two subcategories, with one referring to instances within the lecture and others across lectures in the given course materials. This suggests that one may need to split the PRE tag into two sub-tags in future work.

Speech Acts (SA)

The performance of SA tags follows the normal observation predicted by the frequencies of these tags in the dataset. For this reason, no further inspection is needed in this regards as the model performance are severely impacted by the tag frequencies. For example, SUG in Physics lectures had fewer occurrences in the datasets and the model had very poor performance.

Interaction with Audience (IA)

The MAC tag had a fewer number of occurrences (only 326) compared to other tags, such as INT, MAT, PRE and EMP, but it had higher performance (average F1-score of 68.21) than these other tags, which had an average F1-score in the range of 38.52 to 61.41 across disciplines. In addition, as the number of AAR examples in the dataset for the physics lectures was higher than in the economics lectures, its performance in Physics was better (F1-score of 36.11) than in Economics (F1-score of 4.17). This can again be attributed to the variants problem of these metadiscourse expressions, and the model needs more instances in order to behave properly.

This in-depth analysis for each metadiscourse tag confirms the hypothesis that when modelling metadiscourse tags, the model performance is affected by the frequency estimates of these categories in the dataset. However, for the tags that the model misclassifies even though they have high frequencies (such as MAT, ENU and PRE), one possible reason for such behaviour is that many of the metadiscourse expressions used to indicate these categories were more varied and therefore the model needs more occurrences of these categories to be able to perform well. Another possible reason is related to the way the annotation was

done, in particular for the MAT and PRE tags. For instance, the annotators might have misunderstood the task of these tags even though they reported a high agreement in the range of 68 to 83. This indicates that to further improve the results in future work, one needs to review the annotation guidelines by providing more examples of the task and what should and should not be considered. Another possible solution is to split the instances of the PRE tag category into previewing within the lectures and previewing across lectures, as was suggested above.

6.5 Conclusion

In this chapter, the applicability of classifying metadiscourse tags automatically using a single-layer CNN was investigated. First, an overview of how to use CNNs from an NLP perspective was presented. Then, the adopted CNN architecture was explored, along with detailed descriptions of the features used and how to represent them in continuous space. Further, an in-depth analysis of the best practice for the use of pre-trained word embedding vectors has been provided. An important finding is that fine-tuning the pre-trained embedding vectors for *word2vec* provides better results than using fixed vectors. Also, the combination of both POS and prosodic features, along with the *word2vec* word vectors, improved the results over the baseline configurations of the CNN.

Results show that the use of a CNN outperforms the SVM model by a large margin and that this increase in model performance is also consistent – both per class and per discipline. An observation made in both models was that there is a correlation between high-frequency occurrences and high performance but that this is not the case for some of the tags. To understand this further, an in-depth analysis was conducted on some specific tags that the model was confused about, which confirmed the need to adjust the annotation guidelines.

The approach presented in this chapter led to the successful automatic identification of a set of metadiscourse tags compared to the baseline. As these tags indicate the per sentence function, they could also serve as local indicators of the higher level structure of the discourse. This latter hypothesis is investigated in the next chapter, where we will utilise metadiscourse tags as features, along with others lexical features, for the task of thematic discourse segmentation.

Chapter 7

Exploiting Metadiscourse Tags for Discourse Segmentation

The last four chapters presented an approach for metadiscourse tagging in academic lectures with four stages involving a corpus of metadiscourse in academic lectures from two different disciplines, ASR for academic lectures, features exploring and baseline tagging model using SVMs, and an improved metadiscourse tagging model based on both CBOW and CNNs. The results of this CNN-based metadiscourse tagging model can be applied to boost the performance in several downstream applications such as thematic discourse segmentation, and summarisation.

This chapter describes the thematic discourse segmentation model that has been developed to evaluate the usefulness of the presented approach of metadiscourse tagging in academic lectures. Section 7.1 introduces the thematic discourse segmentation task and how to use these tags to improve the model. Section 7.2 describes the previously introduced discourse segmentation for spoken language. The implementation of the proposed segmentation model is described in Section 7.3. The performance of this model is measured using commonly known metrics – Pk and WindowDiff (WD) – and an analytical study is conducted using several test cases, including the use of ASR outputs, in Section 7.4. A concluding discussion is set out in Section 7.5.

7.1 Introduction

Metadiscourse has proven to be effective in various applications, for example, summarising a meeting according to its activities (Niekrasz, 2012), argumentative zoning for scientific research articles (Teufel and Moens, 2002), and most recently, it will be used in building

presentation skills tools using TED Talks (Correia et al., 2014b). In this work, the particular interest is to investigate the usefulness of metadiscourse tags for the task of the automatic thematic discourse segmentations of academic lectures. The theme of the segment means that it can carry a functional or topical structure and the task is to segment lecture discourse based on these themes. Alharbi and Hain (2015) show the ability of the state-of-the-art discourse segmentation models in segmenting OCW lectures materials. Additionally, the Alharbi and Hain (2015) study reported that the best results among these models were obtained for models that used discourse cues features such as the Bayesian segmentation (BSEG) model proposed by Eisenstein and Barzilay (2008). As mentioned in Chapter 1, such formulation situates this task in the area of discourse segmentation.

Several studies have tackled the task of automatic discourse segmentation for speech communication using a variety of different features (Eisenstein and Barzilay, 2008, Galley et al., 2003, Hearst, 1997, Hsueh et al., 2006, Mohri et al., 2010). Most of these models rely on the concept of lexical cohesion to detect segment boundaries. Lexical cohesion is defined as lexical chains that are related to each other (*e.g.* term repetition), and hence these model hypothesis boundaries when there is a change in the vocabulary. Additionally, early works have addressed the task by utilising certain discourse cues as indicators of segment boundaries such as "now", "so", or "well" (Grosz and Sidner, 1986, Hirschberg and Litman, 1993b). Other studies have used both lexical cohesion and discourse cues in the segmentation model (Eisenstein and Barzilay, 2008, Galley et al., 2003), which show great improvements over previous models, particularly for written discourse, though less so for spoken discourse. This indicates that a knowledge about discourse structure (*i.e* discourse cues) has a positive impact on the model's performance as it may capture some aspects of the rhetorical functions of lecture discourse.

Additionally, the manual annotation of the thematic boundaries reveals that mixed approach of topical/functional is needed when automatically segmenting these lectures, as shown by the discussion in Section 1.1 about the example of Physics and Economics lectures segments, illustrated in Figure 1.1 (A) and (B). This example shows that lecture segments are labelled manually according to both topical and functional content of the lecture. The same observation of a mixed approach to functional/topical segment boundaries for meetings discourse was also noticed by Niekrasz (2012). However, it is a big challenge to represent lecture content automatically in such a strategy as it requires a level of understanding of the lecture content to locate these functional regions in the discourse. This may be required to apply a specific process to the audio transcripts to locate those regions indicated – for example introductions, examples or review areas – through rhetorical functions in the transcripts, without fully understanding the lecture discourse. This supports the hypothesis that metadiscourse tags that reflect the discourse function of the utterances can be of great help to further improve the thematic discourse segmentation model of academic lectures.

To achieve this, the work in this chapter is organised as a sequence of steps towards the objective of utilising metadiscourse tags for thematic discourse segmentation tasks: First, each utterance is represented by several linguistic features including lexical cohesion scores, metadiscourse tags, and discourse cues. Second, investigation is carried out into the appropriateness of the SVM classifier for the thematic discourse segmentation task, which is referred to as TDS-SVM. Third, the impact of the model performance on ASR outputs is also investigated, in order to validate the robustness of the developed model. Fourth, the obtained results of the TDS-SVM model are compared with existing state-of-the-art discourse segmentation models. The following sections review work in developing segmentation models for spoken discourse, and then provide some details about the proposed approach for thematic discourse segmentation, along with how the evaluation is accomplished, and results analysis on several test cases including the application on ASR outputs, and end with a concluding discussion.

7.2 Related Work

This section reviews the studies of thematic discourse segmentation models, as either supervised or unsupervised models for spoken discourse.

7.2.1 Unsupervised Models

In the last decade, several studies have been proposed for unsupervised segmentation of text. Most of these models are based on the concept of lexical cohesion. However, few studies tried also to involve other elements, such as discourse or acoustic cues. This section, therefore, is organised according to the way lexical cohesion is modelled either as similarity-based, language model-based or topic model-based.

The *TextTiling* model developed by Hearst (1997) and *C99* model presented by Choi (2000) treat a document as a series of blocks, where each block is composed of several sentences. They then measured cosine-similarity between each consecutive block of words. A segment boundary was signalled when there was a drop in the cosine score between two adjacent blocks. Hearst (1997) measured the similarity solely based on word frequency, while Choi (2000) used divisive clustering with a matrix-ranking scheme. Another unsupervised similarity-based model is the LCseg proposed by Galley et al. (2003). LCseg models lexical chain repetitions of a given lexical term throughout a fixed-length window of sentences and then chooses segmentation points at the local maxima of the cohesion function. Along this line of research, Malioutov and Barzilay (2006) presented a model that aimed to optimise

the normalised minimum-cut criteria based on a variation of the cosine similarity between utterances.

An earlier model based on language model was presented by Utiyama and Isahara (2001) which attempted to find segmentation boundaries with compact language models. In a similar fashion, Eisenstein and Barzilay (2008) introduced a segmentation model based on generative Bayesian model in which each sentence is modelled by a language model related to a segment. Then it computes the maximum likelihood estimates by looking at the entire sequence of sentences at specific segment boundaries. Further, it uses the initial of the potential boundary utterances as discourse cues for the unsupervised model, which is an extension of the work by Galley et al. (2003), who automatically identified discourse cues using manually labelled boundaries in a supervised fashion, as discussed in the following section. The Eisenstein and Barzilay (2008) study has shown a positive effect on the segmentation model when they used discourse features along with the semantic ones. This is not surprising as the use of discourse cues was predicted to occur at intentional segment boundaries (Grosz and Sidner, 1986).

Several studies have proposed unsupervised models based on Latent Dirichlet Allocation (LDA; Blei et al. (2003)). LDA is a generative model which uses latent structures to model the underlying similarities among observations and it is widely adopted in text analysis to model the shared topics among documents. Topic model-based segmentation was first introduced by Sun et al. (2008) and built upon by Misra et al. (2009). The most recent LDA based segmenter is *TopicTiling* (Riedl and Biemann, 2012), which undertakes linear topic segmentation with a pre-trained LDA topic model and estimates the similarity between segments to evaluate text coherence based on a topic vector representation with cosine similarity. The most common topic ID is given to each word in the sentence using Gibbs' sampling in order to maintain efficiency. Du et al. (2013) have shown a hierarchical Bayesian model, which jointly models Bayesian segmentation and structured topic modelling STM. The model provides remarkable performance over various models in both written and spoken texts.

7.2.2 Supervised Models

As the case with unsupervised models, existing supervised approaches to spoken discourse segmentation are based on the concept of lexical-cohesion via semantic similarity, but other features such as discourse cues and acoustic cues are also included, and hence a supervised approach was needed to combine these knowledge sources. For instance, Litman and Passonneau (1995) and Passonneau and Litman (1997) used decision trees to segment spoken monologues known as the pear stories, where each candidate boundary is classified as either boundary or not. The model represents the utterance using a set of linguistic features, including referential noun phrases to represent the concept of lexical cohesion and other acoustic

features such as prosodic cues. The inclusion of discourse cues is based on the observation that there is a correlation between such cues and discourse segment boundaries. Their extraction procedure is drawn from an empirically selected list of words, some of which are similar to the discourse cues employed in this work.

Galley et al. (2003) proposed a similar feature-based model to segment spoken discourse, in particular multi-party speech, that is, meetings. The discourse segmentation task was defined as a binary classification problem classifying each sentence break as boundary or no boundary. Additionally, the model used multiple features such as lexical, discourse and prosodic cues. The lexical features were modelled based on lexical cohesion criteria. More precisely, Galley et al. (2003) used the LCseg model to compute the lexical cohesion score in unsupervised fashion (this model is discussed further in the next section). These features of lexical cohesion scores (the posterior that computed by the LCseg and the raw lexical cohesion scores), and discourse and prosodic cues are then combined in the segmentation model in a manner similar to the knowledge source combination proposed by Beeferman et al. (1999) and Tür et al. (2001), who also used the output of lexical-based model such as HMM as input to an overall discourse segmentation classifier. A follow-up study by Hsueh et al. (2006) used the same approach to investigate the impact on the model performance on meeting discourse by investigating its effectiveness in two further cases: subtopic boundaries and on ASR outputs. Results showed that the model performance degraded on both subtopic boundaries and ASR outputs compared to topic boundaries when using the same feature combinations.

In addition, Georgescul et al. (2006) and Georgescul et al. (2007) investigated the suitability of using SVMs for the task of meeting segmentation. The input features in both studies are word distributions in the form of bag-of-words to represent each utterance as multiset of words, disregarding word order and grammar, a representation commonly used in both NLP and information retrieval (IR). Georgescul et al. (2007) explore additional features for the approach, such as prosodic cues, syntactic features and discourse cues. The main intuition in their work is to check the appropriateness of the SVM approach in combining prosodic features with high-dimensional lexical features. Results indicate remarkable improvements over existing SVM-based models on the same task of meeting segmentation.

Most of the previous models are based on discriminative models to combine multiple features including the lexical cohesion and discourse cues. Kokhlikyan et al. (2013), on the other hand, utilised a generative model in the form of HMM for better lecture segmentation and summarisation. The authors found some phrases that are similar in definition to the metadiscourse phenomena. Therefore, the authors proposed an extraction method that aimed to extract these phrases and then use them in the automatic segmentation model based on pedagogical lecture elements. Such pedagogical elements allow them to segment the lecture

content into a fixed structure such as *Introduction*, *Background*, *Main Topic*, *Question*, and *Conclusion*. This is in contrast to the nature of the lecture discourse which may or may not have such structure.

The model presented in this chapter is similar in spirit to the supervised feature-based studies introduced by Galley et al. (2003) and Hsueh et al. (2006) in using the lexical cohesion scores, but the current work differs in three main respects. First, this work investigates the usefulness of SVMs for the task of lecture segmentation similar to Georgescul et al. (2006) and Georgescul et al. (2007). Second, this work proves the effectiveness of the metadiscourse tags as discourse features, which was not investigated before for the discourse segmentation task. Finally, as it is hard to replicate the supervised experiments of these studies, the state-of-the-art unsupervised discourse segmentation models are used to segment the OCW lectures for the purpose of analysis and comparison, as demonstrated in Section 7.4.

7.3 Metadiscourse for Lecture Segmentation

The aim of this work is to investigate whether metadiscourse tags are an effective features for the thematic segmentation task of academic lectures. In this framework such task is considered as a binary classification problem in which each utterance break should be classified as a boundary or not. In studying the effect of metadiscourse tags on this classification problem, there are other features that were used as well as explained in section 7.3.1. Note that the thematic discourse segmentation task employed SVM classifier as described in section 7.3.2, and the input to the TDS-SVM model is a vector representation of the utterance boundary to be classified and its context. Each dimension of this input vector contains feature that characterises the utterance.

7.3.1 Features

Lexical Cohesion

The lexical cohesion features for each utterance break were obtained by utilising the unsupervised LCseg model developed by Galley et al. (2003), as was briefly explained in Section 7.2. LCseg uses lexical chains of word repetitions in computing the lexical cohesion features, instead of word counts between two contiguous windows as the case with TextTiling. A lexical chain is comprised of all repetitions of a term from its first to the last occurrence in the discourse (Galley et al., 2003). The LCseg model hypothesis is that a major topic shift is predicated to occur when there are strong term repetitions. The input to the model passes through several steps in similar fashion to other segmentation models such as TextTiling

(Hearst, 1997). These steps are tokenisation, stop word removal and finally stemming of the remaining words to keep terms that are semantically similar close together. The final step is performed using an extension of Porter's stemming algorithm (Xu and Croft, 1998). For each utterance break u_i, the LCseg produces two lexical cohesion scores. The raw lexical cohesion value (LCV) is computed by measuring the similarity of lexical chains between two contiguous analysis windows of fixed size W_{LC} using cosine similarity. The window size is determined based on experimental trials, as shown in Table 7.3 in section 7.4.1. The other score is the probability of topic shift (denoted by LCP) indicated by the sharpness of change in the raw lexical cohesion score. These two scores are then fed into TDS-SVM as a representative of the lexical cohesion features for each utterance break.

Metadiscourse Tags

The metadiscourse tags are used as features in the lecture segmentation model in two forms: specific tags and generic tags. In the specific tags case, the number of any metadiscourse tags within a specified window before and after the candidate thematic boundary is counted. This component is formalised as follows. Let INT_i be the number of the metadiscourse tag INT in an utterance u_i. Then, $\vec{ul}_i^{INT}$ represents the number of INT that occur in utterances situated before u_i:

$$\vec{ul}_i^{INT} = (INT_{i-W_{INT}+1}, INT_{i-W_{INT}+2}, ..., INT_i),$$

Similarly, the number of INT tag occurring in utterances situated after u_i in an interval of size W_{INT}:

$$\vec{ur}_i^{INT} = (INT_{i+1}, INT_{i+2}, ..., INT_{i+W_{INT}}),$$

where W_{MD} refers to the window size of metadiscourse tags in general, which will be determined based on experimental trials in Table 7.3 in section 7.4.1. Then a normalisation process is applied on the vectors $\vec{ul}_i^{INT}$ and $\vec{ur}_i^{INT}$, by dividing each value in the vector by the sum of all entries in the vector. The total number of MD that occurred around the utterance break u_i can then be computed as $\vec{u}_i^{INT} = \vec{ul}_i^{INT} + \vec{ur}_i^{INT}$. The other 18 metadiscourse specific tags are handled in the same manner.

In the generic tags case, the same process is done but this time for only four tags: ML, DO, SA and IA. Thus, let DO_i be the number of the general tag that occurred around u_i. Then $\vec{ul}_i^{DO}$ contains the number of DO tag that occur in utterances situated before u_i in an

interval of size W_{DO}:

$$\vec{ul}_i^{DO} = (DO_{i-W_{DO}+1}, DO_{i-W_{DO}+2}, ..., DO_i),$$

In a similar fashion, the other vectors $\vec{ul}_i^{ML}$, $\vec{ul}_i^{SA}$, $\vec{ul}_i^{IA}$ contain the number of metadiscourse tags of the other generic tags: ML, SA, and IA, respectively occurring before u_i. Then, the vectors $\vec{ur}_i^{ML}$, $\vec{ur}_i^{SA}$, $\vec{ur}_i^{IA}$ will contain the number of ML, SA and IA, respectively occurring after u_i. Finally, the $\vec{u}_i^{ML}$, $\vec{u}_i^{DO}$, $\vec{u}_i^{SA}$ will contain the total number of each generic tag around the utterance u_i after applying a normalisation process as in the case with specific metadiscourse tags.

Discourse Cues

Several studies have shown the effectiveness of discourse cues in boosting the thematic discourse segmentation model performance (Galley et al., 2003, Hsueh et al., 2006, Litman and Passonneau, 1995, Passonneau and Litman, 1997). Therefore, in this work, the use of discourse cues as features in the segmentation model is also considered to analyse its impact on the model performance compared to metadiscourse tags. The correlation between each word in the lecture and manually labelled thematic boundaries are computed. The final list of discourse cues in both disciplines is obtained by selecting words that are statistically correlated with thematic boundaries.

More formally, for every word in the lectures corpus, its number of occurrences is counted near any thematic boundary against those far away. The chi-squared test computes the significance of the near-against distinct-statistics by comparing it with the overall statistics, where the null hypothesis is assumed. The terms whose χ^2 rejected the hypothesis under a 0.01-level confidence (the rejection criterion is $\chi^2 \geqslant 6.635$), are selected. In Table 7.1 the counts of the word "today" are listed and also the overall counts in both Physics and Economics lectures. The computed χ^2 values are 68.87 and 150.78 in the Physics and Economics corpus, respectively. This means that "today" is indicative of the presence of a thematic boundary in both disciplines. Based on this strategy, the final set of words that were considered as discourse cues in this study are listed in Table 7.2. By examining this list we noticed that there are some words that are not considered usually as discourse markers, such as 'and' or 'about'. The main reason for including such words in the list is because the strategy used to extract these words allows us to consider any word in the utterance as a candidate discourse marker and does not restrict that option only to initial utterances, while other words, such as 'We' or 'Today', in the list follow the normal expectation as discourse markers as they occur as initial utterances.

	Near	Distant	Near	Distant
today	65	83	128	164
Other	85939	402211	70986	422410

TABLE 7.1: $\chi^2 = 68.87$ in Physics and 150.78 in Economics.

Physics		Economics	
DC	χ^2	DC	χ^2
today	68.87	**about**	477.95
topic	63.17	**talk**	375.38
last*	59.13	lecture	179.15
talk	36.97	**today**	150.78
we	30.05	**topic**	85.78
okay*	32.46	they	68.11
so*	31.97	**last***	52.38
let's	30.51	you	51.01
about	29.88	**we**	46.35
have	25.5	and*	42.35

TABLE 7.2: Automatically selected discourse markers which are significant according to the chi-squared value at the level of $p < 0.01$. Boldface indicates that these markers are common across the two disciplines. Asterisks indicate discourse markers that been described by previous works (Hirschberg and Litman, 1993a)

Finally, each discourse cue in the selected list is automatically examined if it occurs within a fixed size window before and after an utterance u_i break. In a manner similar to the case with metadiscourse tags described previously, the number of discourse cues occurring in utterances situated before u_i in an interval of size W_{DC}:

$$\vec{ul}_i^{DC} = (DC_{i-W_{DC}+1}, DC_{i-W_{DC}+2}, ..., DC_i),$$

where DC_x is the number of discourse cues in utterance u_i. Similarly, the number of discourse cues occurring in utterances situated after u_i in an interval of size W_{DC}:

$$\vec{ur}_i^{DC} = (DC_{i+1}, DC_{i+2}, ..., DC_{i+W_{DC}}).$$

Then the total number of DC that occurred around the utterance u_i will be computed as $\vec{u}_i^{DC} = \vec{ul}_i^{DC} + \vec{ur}_i^{DC}$.

7.3.2 Model

As stated before, the thematic discourse segmentation is defined as a binary classification task using SVMs. An SVM is a discriminative classifier and can learn from small training examples, and shows effectiveness in a number of sentence-level classification tasks including

the metadiscourse tagging, as discussed in Chapter 5. Additionally, it shows high performance on previous thematic segmentation task in both written (Kauchak and Chen, 2005) and spoken discourse, specifically meetings (Georgescul et al., 2007). However, speech exhibits a large degree of variability, disfluencies and speakers' styles, which add another layer of complexity to the model compared to written discourse. In this work, SVMs are employed in the same manner as in Georgescul et al. (2007) but this work will investigate their effectiveness for academic lectures. In the following, some highlights are given on how SVMs are used for thematic lecture segmentation in a supervised fashion.

Under this supervised learning paradigm, a training set in which each candidate thematic boundary is labelled as either "boundary" (1) or "non-boundary" (-1) is used to train a classifier to predict whether each new example in the test set belongs to the class -1 or 1. The objective here is to determine the advantage of integrating metadiscourse tags in its two forms: generic tags and specific tags, alongside other features to improve automatic thematic segmentation in academic lectures, as well as in the case when ASR transcripts are used instead.

The support vector learner $\mathcal{L}$ is given a training set of n samples, which are referred as $S_{train} = ((\vec{x}_1, y_1), ..., (\vec{x}_n, y_n)) \subseteq (X \times Y)^n$ (Georgescul et al., 2007). Every training sample is represented by a vector $\vec{x}$ which contains the set of all features described in Section 7.3.1, and the class label y in which $y \in \{-1, 1\}$. The hypothesis function is similar to Equation 5.3.2, and has the form:

$$f(\vec{x}) = sign(\vec{w}, \vec{x} + b),$$

where $f : X \to \{-1, 1\}$, the support vector learner $\mathcal{L}$ then attempts to learn the decision function f using the training set S_{train} by reducing the errors using the structural risk minimisation principle (Vapnik, 1995). The parameters $\vec{w}$ and b are optimised with Lagrangian theory. In this work, the radial basis functions are used as a kernel, as provided by the Scikit-learn library (Pedregosa et al., 2011).

7.4 Experiments and Results

7.4.1 Experimental Setup

Parameter Estimation

As discussed in Section 7.3.1, there is a need to determine the optimal window size for each feature type. To achieve that, a search process was performed using the TDS-SVM model

Type	Tag	Window Size (# U)
Lexical Cohesion	LC	4
Metadiscourse Tag	MD	10
Discourse Cues	DC	5

TABLE 7.3: Parameters for feature analysis. U denotes utterances.

to analyse features in a window preceding and following each utterance u_i in order to tune the window size on the given lecture's corpus. The value that provided the best performance was selected and it is listed in the third column of Table 7.3 for each feature type. It is worth noting that in using the LCseg model, other parameters are needed to determine the hiatus h, which is the length for dividing a chain into parts. In this work, h has been set to 11 for both disciplines, Physics and Economics. There are other parameters that are needed in setting the LCseg, such as α which is set to $\frac{1}{2}$, and it is used for thresholding limits for the hypothesised boundaries. Note that the selected window size for metadiscourse tags is applied in the same manner on the two metadiscourse forms: generic and specific.

Dataset Used

The thematic segmentation model is evaluated using the OCW lectures corpus which is described in Table 2.4 in Section 2.4. As stated previously, the manual thematic segmentations are labelled by the teaching staff at MIT and Yale universities and these manual segmentations show that the lecture is distinguished by a high-level structure. These labelled segment boundaries are used as the ground truth reference to evaluate the TDS-SVM model performance. To obtain the ground truth on ASR outputs, the procedure described in Section 5.4.1 is also followed here, but for transferring the thematic boundaries at utterance level.

The evaluation process was carried out by performing 49-fold and 57-fold cross-validation for economics and physics, respectively. The number of folds in both disciplines was selected based on the number of lectures in each discipline. Each model is trained on 48 economics and 56 physics lectures and test on the one held-out lecture.

Evaluation Metrics

All experiments are evaluated with regards to the widely used error metrics P_k (Beeferman et al., 1999) and *WindowDiff* (WD) (Pevzner and Hearst, 2002). Both metrics run a window throughout a document, and evaluate if the utterances on the edges of the window were suitably segmented with regards to one another. In other words, both metrics are window-based in which a sliding window is passed over the utterances and its agreement with the

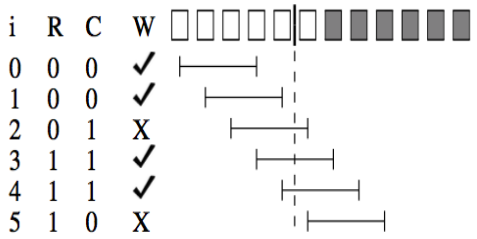


FIGURE 7.1: Illustration of counting boundaries in windows, the figure adopted from (Scaiano and Inkpen, 2012). Each rectangle represents an utterance, while the shade indicates true segments (reference segmentation). The vertical line represents the hypothesis boundary and the window size is 5. The columns i, R, C, W represent the window position, the number of boundaries from the reference (true) segmentation in the window, the number of boundaries from the computed segmentation in the window, and whether the values agree, respectively.

reference segmentation is counted. Typically, windows size k is chosen to be half the average segment length as follows:

$$k = \frac{N}{2 * number\,of\,segments},$$

where N is the total number of utterances.

P_k is defined as the probability of segmentation error. To calculate P_k, a window of fixed width k is selected and moved across the document; in each segmentation point, it is checked whether the predicted segmentation point is correct when compared with the reference segmentation (Purver, 2011). The error function for P_k is:

$$P_k = \frac{1}{N-k} \sum_{i=0}^{N-k} (R_{i,i+k} \otimes C_{i,i+k}) \tag{7.1}$$

where $\otimes$ is the XOR operator – *i.e.* "both or neither" (Purver, 2011). N denotes the number of utterances, R denotes the number of boundaries from the reference in the specified reference W, and C represents the number of boundaries from the computed segmentation in W.

WD, on the other hand, works in a similar fashion by moving a fixed-width window of size k across the lecture document, but in this metric, windows are considered 'correct" only if they assign the same number of segment boundaries between their start and end (Purver, 2011, Scaiano and Inkpen, 2012). Figure 7.1 illustrates the process of computing WD. The basic error function for a window is then computed as:

$$WindowDiff = \frac{1}{N-k} \sum_{i=0}^{N-k} (|R_{i,i+k} - C_{i,i+k}| > 0) \tag{7.2}$$

	Physics		Economics		Average	
Features	P_k	WD	P_k	WD	P_k	WD
LC	35.70	36.01	33.43	35.88	19.57	35.95
LC+DC	32.54	34.25	31.18	33.76	31.86	34.01
LC+MD_{All}	29.01	30.47	27.81	28.84	28.41	29.66
LC+DC+MD_{All}	**27.91**	**30.00**	**24.38**	**26.43**	26.15	28.22

TABLE 7.4: Results of the TDS-SVM model on the reference transcripts for the set of experiment 1 using specific metadiscourse tags: LC denotes lexical cohesion, MD denotes using all specific metadiscourse tags and DC denotes discourse cues. Bold-faced values are scores that are statistically significant.

Equations 7.1 and 7.2 indicate that WD is stricter than P_k as it needs the number of interfering segments between each two sentences to be exactly the same in both the predicated (C) and reference (R) segmentations, while P_k only checks if the two sentences are in the same segment. P_k and WD are penalties, so lower values show superior performance. The evaluation source codes for these two metrics are provided by Malioutov and Barzilay (2006).

7.4.2 Results

This section presents the results of the TDS-SVM model of feature combinations, including metadiscourse tags on both reference and automatic transcripts. It also presents a comparison between the results of manually annotated metadiscourse tags and the ones obtained automatically on both reference and ASR transcripts. Note that all the 19 specific metadiscourse tags were used in this set of experiments. However, another set of experiments is performed to compare the effects of generic and specific metadiscourse tags on the performance of the TDS-SVM model. Finally, the settings of the feature-based TDS-SVM that provides the best results are used to compare with state-of-the-art unsupervised lexical cohesion-based discourse segmentation models.

Features Combinations using Reference Transcripts

At first, the impact of the metadiscourse tags is determined on the TDS-SVM segmentation performance by comparing it with other features, namely lexical cohesion and discourse cues. The hypothesis in this set of experiments is that considering metadiscourse contributes to the effectiveness of the model. The TDS-SVM model was tested using a range of features to investigate how the performance is improved. Then, these features were combined together, resulting in the combination of lexical cohesion (LC), metadiscourse (MD) and discourse cues (DC). The reliance on LC scores as the main features in this work comes from the fact that previous works have proved the effectiveness of such features in finding the segment boundary. Note that all the specific metadiscourse tags were used in this set of experiments.

	Physics		Economics		Average	
Features	P_k	WD	P_k	WD	P_k	WD
LC	38.93	40.65	36.84	38.19	37.89	39.42
LC+DC	38.11	40.19	36.27	37.96	37.19	39.08
LC+MD_{All}	34.88	36.76	31.22	33.57	33.05	35.17
LC+DC+MD_{All}	**34.82**	**36.35**	**30.87**	**33.10**	32.85	34.73

TABLE 7.5: Results of the TDS-SVM model on the ASR transcripts for the set of experiment 1 using specific metadiscourse tags: LC denotes lexical cohesion, MD denotes using all specific all metadiscourse tags and DC denotes discourse cues. Bold-faced values are scores that are statistically significant.

The results in Table 7.4 confirm the hypothesis–using metadiscourse tags helps the segmentation model across the two disciplines. For example, on manually transcribed lectures, the TDS-SVM model yields 35.70 P_k measure on Physics and 33.43 on Economics when using only lexical cohesion features. The inclusion of discourse cues improves the model performance, and this is consistent in the two disciplines. Such improvements were also noticed in similar studies on discourse segmentation reported by Galley et al. (2003) and Hsueh et al. (2006), but for meetings speech. As expected, the introduction of MD specific tags improved the model performance more than the improvements that were obtained from adding DC. As Table 7.4 shows, the improvement is consistent across all lecture datasets. However, the impact on Economics lectures when using such features was more than on Physics lectures. These findings are in line with what was obtained with the metadiscourse tagging task presented in Chapters 5 and 6– *i.e.* the results for Economics lectures are better than those for Physics. This can be attributed to the fact that Economics lecturers often use slides in their talk, which may make their discourse more organised compared to Physics lectures. Additionally, by combining all features, the TDS-SVM model is statistically significantly improved across both disciplines by large margin as the average P_k measure is 26.15 compared to 28.41 when only using both LC and MD tags, and 31.86 when both LC and DC are used. This is expected as both DC and MD have shown remarkable improvements in the model performance over the use of LC individually.

Features Combinations using Automatic Transcripts

To investigate the TDS-SVM model's robustness on error-full transcripts such as ASR outputs, the same set of features that were used before is investigated here as well. However, these extracted features from ASR outputs are different than those extracted from reference transcripts for two reasons: 1. The errors from ASR model are propagated when extracted the lexical cohesion features. 2. These errors also affect the process of extracting the set of discourse cues. Note that the metadiscourse features are extracted here based on the manual

	Physics		Economics		Average	
Models	P_k	WD	P_k	WD	P_k	WD
	REF					
LC+DC+$MD_{All_{man}}$	**27.91**	**30.00**	**24.38**	**26.43**	26.15	28.22
LC+DC+$MD_{All_{auto}}$	30.44	33.26	26.39	29.07	28.42	31.17
	ASR					
LC+DC+MD_{All}	**34.82**	**36.35**	**30.87**	**33.10**	32.85	34.73
LC+DC+$MD_{All_{auto}}$	36.75	39.18	32.91	34.09	34.83	36.64

TABLE 7.6: Results of the comparison between TDS-SVM and using automatically detected metadiscourse tags models on automatic transcripts. Bold-faced values are scores that are statistically significant.

annotation of the metadiscourse tags, and projected to the ASR outputs via aligning process as described in Section 5.4.1.

The results in Table 7.5 show a remarkable degradation in the model performance when ASR transcripts are used, compared to the case when reference transcripts are used. In addition, it seems that the use of discourse cues does not play a significant role in improving the model performance compared to the use of lexical cohesion scores, and this is noticed in both disciplines. However, the addition of all specific tags of metadiscourse significantly improved the model performance compared to the case when discourse cues are used. The combination of all features outperform the use of both lexical cohesion and metadiscourse tags. This can be attributed to the fact that the metadiscourse tags used here are the ground truth annotations which were obtained on ASR outputs by alignment process as discussed previously. This suggests that there is a need to investigate the model performance when automatically detected metadiscourse tags are used, as will be explained in the following section.

Manual vs. Automatic Metadiscourse Tags

To have further insight on the effectiveness of using metadiscourse tags, Table 7.6 shows the performance of the TDS-SVM segmentation model when automatically identified metadiscourse tags were used in both reference and ASR transcripts. The automatic identification of metadiscourse tags was obtained from the best metadiscourse tagging results of the MDT-CNN in Chapter 6. The previously best obtained results for either reference or ASR transcripts from Table 7.4 and Table 7.5 features combinations of the TDS-SVM segmentation model, also included here in order to have a fair comparison. As expected, the results reveal that there is a marked decline of about 2–3% in the model performance compared to the use of the manually annotated metadiscourse tags in both types of transcripts, and also in both disciplines. Nevertheless, the decline for Economics lectures is less than it is for

	Physics		Economics		Average	
Features	P_k	WD	P_k	WD	P_k	WD
LC+DC+MD_{All}	**27.91**	**30.00**	**24.38**	**26.43**	26.15	28.22
LC+DC+MD_{MC}	33.86	35.07	30.54	31.14	32.20	33.11
LC+DC+MD_{DO}	28.29	30.95	25.77	27.32	27.03	29.14
LC+DC+MD_{SA}	29.31	32.39	26.12	28.41	27.72	30.40
LC+DC+MD_{IA}	31.62	33.08	29.89	31.69	30.76	32.39

TABLE 7.7: Results of the TDS-SVM model on the reference transcripts for the set of experiment 1 using specific metadiscourse tags: LC denotes lexical cohesion, *MD** denotes using all metadiscourse tags and DC denotes discourse cues. Bold-faced values are scores that are statistically significant.

Physics. Again, this may due to the fact that the metadiscourse tagging model in particular MDT-CNNs performed better for Economics than Physics, as explained in Section 6.4.

Generic vs. Specific Metadiscourse Tags

Table 7.7 shows the results of using each of the four generic metadiscourse tags individually in the TDS-SVM segmentation model. The first row shows the best results using specific metadiscourse tags from the previous Table 7.4. It is included here to provide a comparison with the results of using generic tags. In general, the results indicate that there is a correlation between frequently occurring metadiscourse tags and the improvements in the performance of the TDS-SVM model. When Discourse Organisation (DO) tags are considered, the model performance improved compared to the other three generic tags as the P_k measure is 28.29 and 25.77 in Physics and Economics, respectively. Similarly, Speech Acts (SA) tags improved the performance as well, but not as much as DO tags. This is expected since the specific tags that belong to DO are more inductive of segment boundaries. The lecturers will occasionally use DO tags such as *Introduction* (INT) and *Conclusion* (CON) to mark the beginning and the ending of a segment, whereas the SA tags such as *Emphasising* (EMP) and *Exemplifying* (EXE) may occur around segment boundaries but not as frequently as the DO tags. Nevertheless, the obtained results when using all specific tags, referred as MD_{All}, are significantly better than using either DO or SA tags alone.

The other generic metadiscourse tags such as *Metalinguistics Comments* (MC) and *Interaction with Audience* (IA) perform less effectively than DO and SA. This may be due to the fact that these tags are generally less frequent in the gold datasets than other tags. For instance, both MC and IA have 2254 occurrences compared to 7939 tags from both DO and SA, as indicated in Table 5.2 in Section 5.4.1. However, the use of IA provides better results (P_k = 31.62 in Physics and 29.89 in Economics) than the contribution of MC tag (P_k = 33.86 in Physics and 30.54 in Economics) to the segmentation model. This may be because the IA tags often use such metadiscourse expressions to interact with students. For

example, speakers use this to ensure they are in line with what was introduced in the lecture by asking them if there are any questions about what had been introduced so far in the lecture, which often occurs at the end of the discourse segment and the beginning of a new segment. This indicates that although MC tag are more frequent than IA, it seems that they may not generally occur around segment boundaries as lecturers may use the MC tag to clarify some points using the specific tag of *Clarifying* (CLA) or to comment on the use of some terminology within the segment using the *Comment on Language Form* (CLF) tag.

7.4.3 Comparison with the State-of-the-Art

The best obtained results from previous settings of TDS-SVM model were compared with state-of-the-art lexical-cohesion based segmentation models. This set of experiments also allows us to establish how well the metadiscourse features perform in relation to current state-of-the-art models. However, comparing the performance of TDS-SVM model to other similar existing supervised model is not straightforward due to differences in corpora, in experimental design, and/or different input features. For this reason, state-of-the-art unsupervised discourse segmentation models are used instead, and evaluated on the same dataset.

The publicly available implementations of these models were used, and optimised in the same settings as specified in the models' papers. These models are: UI (Utiyama and Isahara, 2001), LCseg (Galley et al., 2003), MCS (Malioutov and Barzilay, 2006), and BCseg (Eisenstein and Barzilay, 2008). These models were discussed in detail in Section 7.2. It is worth noting that the LCseg model is the one used to produce the lexical cohesion scores for the TDS-SVM model. The configuration of the LCseg model is the same as the one described in Section 7.3.1 where we used $w = 4$ and $h = 11$ for both disciplines of Physics and Economics. For the MCS, the same configuration reported in Malioutov and Barzilay (2006) was used. The remainder of the models do not need to specify configurations. Additionally, the BCseg model was used in two settings, one based on only lexical cohesion features and the other on both lexical cohesion and discourse cues which were extracted in the same manner as described in Section 7.3.1, but in unsupervised fashion. Note that in these unsupervised models the number of segments were known beforehand.

Table 7.8 shows that the TDS-SVM model significantly outperforms the state-of-the-art models. This gain can be attributed to the use of metadiscourse tags as features in the TDS-SVM model which act as indicators of high-level structures in spontaneous speech such as lecture speech. Lecturers often used the metadiscourse tags to direct students from one concept to another in the given lecture. Moreover, the results indicate that the performance of probabilistic-based unsupervised models such as UI and BCseg give better results ($P_k \in$[35-39]) than other similarly-based models ($P_k \in$[38-41]) such as LCseg and MCS, and this was

	Physics		Economics		Average	
Models	P_k	WD	P_k	WD	P_k	WD
LC+DC+$MD_{All_{man}}$	**27.91**	**30.00**	**24.38**	**26.43**	26.15	28.22
UI	39.49	43.58	36.25	39.79	37.87	41.69
LCseg	40.11	45.20	38.98	41.60	39.55	43.40
MCS	41.33	47.56	39.42	44.68	40.38	46.12
BCseg	39.04	40.25	38.86	40.14	38.95	40.19
BCseg+DC	36.44	40.13	35.07	39.23	35.76	39.68

TABLE 7.8: Results of the comparison between TDS-SVM and other state-of-the-art segmentation models on reference transcripts. Bold-faced values are scores that are statistically significant.

observed in both disciplines. In particular, the use of BCseg with discourse cues provides favourable performance on both models compared to other state-of-the-art models. This is expected as the case with the TDS-SVM supervised model described previously gives better results when discourse cues are used over the reliance on only lexical cohesion features. This slight increase in the BCseg+DC can be attributed to the fact that discourse cues may be used to indicate thematic transitions but not as effectively as metadiscourse tags. In summary, enriching the transcripts with different metadiscourse categories seem to be very effective, even with the use of automatically detected metadiscourse tags, as it gives better results than most of the state-of-the-art models. This is because these tags are more informative about the segment types than the use of the primitive discourse cues. This suggests that the use of metadiscourse tags can be useful as well for the task of labelling these segments, something which will be considered as future work.

7.5 Conclusion

The previous four chapters presented a metadiscourse tagging approach in academic lectures with four-stages including a corpus of metadiscourse in academic lectures from two different disciplines, ASR for academic lectures, features of exploring and baseline tagging model using SVMs, and an improved metadiscourse tagging model based on both CBOW and CNNs. During the development, the first stage was evaluated using inter-annotator agreement, and the second stage evaluated the generated ASR output using the WER. The third and fourth stages evaluated the tagging model using the commonly known metrics of precision, recall and $F1$-score . Such tags show promise for a number of discourse-based applications such as summarisation, meeting segmentation, and argumentative zoning.

In this chapter, we evaluated the usefulness of the identified metadiscourse tags either automatically via the developed approach or manually by developing a model for thematic discourse segmentation for academic lectures. The theme of the segment in such data are

mixed with both topical and functional structures. The model combines both metadiscourse tags that signalled more of the functional structure of the lectures along with the computed lexical cohesion score that reflects the topical structure of the segment. Experiment results on OCW lectures show great improvements in the model performance compared to state-of-the-art unsupervised segmentation models on the same task. Most of these models are based only on the lexical cohesion criteria.

Chapter 8

Conclusion

Metadiscourse tagging is in demand for various discourse analysis applications, such as thematic discourse segmentation that attempts to find structure of the discourse from different domains, the labelling of these discourse segments that can be used for web-browsing, and summarisation. The metadiscourse tagging task seeks to capture knowledge about the discourse structure of academic lectures. It is a fundamental task that helps in identifying the rhetorical structure of lectures discourse that cannot be captured by lexical-cohesion-based approaches. Deciding on a formal definition of the metadiscourse scheme is the key step for discourse analysis applications, including thematic lecture segmentation. In lecture data some utterance sequences are semantically tied to presenting a topical segment, while other sequences represent a functional segment that can be captured by metadiscourse tags. In other words, discourse segments of academic lectures represent a mixed approach of topical and functional structure. For this reason, an analysis of lecture discourse was presented in this thesis by characterising utterances according to the pre-defined metadiscourse tags. The problem of assigning utterances with metadiscourse tags is formulated as a multiclass classification task using both textual and acoustic features.

This thesis is concerned with the metadiscourse tagging of academic lectures from different disciplines, which can be used as discourse features to improve the task of thematic discourse segmentation of academic lectures. For this purpose, a four-stage approach was developed for metadiscourse tagging in academic lectures. Firstly, a corpus of metadiscourse for academic lectures from Physics and Economics courses has been built using an existing scheme that represents functionally-oriented metadiscourse categories (see Chapter 3). The second task is aimed at building an automatic speech recognition system specifically to produce automatic transcripts for OCW materials to alleviate the need for manual transcripts which cost time and effort. However, the reference transcripts lack the time-stamp information needed to be able to evaluate the ASR system (see Chapter 4). For this reason, an alignment system has

been applied to the reference transcripts in order to provide such time information. Then, a strategy was followed to add the metadiscourse annotation onto the ASR outputs. Finally, a model was developed to classify metadiscourse tags on either the reference or ASR outputs, using both textual and acoustic features (see Chapter 5). However, to further improve the quality of the tagging task, another model has been developed that utilises both continuous features representations and CNNs (see Chapter 6). To verify the usefulness of identified metadiscourse tags, a model of thematic discourse segmentation was developed that used such tags as features (see Chapter 7).

The main motivation of this research come from the fact that several universities have started to launch online educational course materials, such as the OCW initiative, to enable free learning under creative common licences. Such efforts give rise to open research questions related to content management and organisation required to serve student information needs. Most of these online platforms rely on labour-intensive processes that are very costly and prone to errors, due to the manual effort required. There is a need to develop a systemic way to unify the content organisation and management processes, to enable linking across platforms. To solve such a problem, many algorithms have been proposed as the starting point for organising such materials, by segmenting lecture content according to subtopics. However, a segment in a lecture is not just represented by subtopic: it can also be represented by the rhetorical functions in the discourse (*e.g.,* introduction, exemplifying, *etc*). Hence, capturing such knowledge about lecture discourse structures and enriching the transcripts with such instances of metadiscourse functions can be of practical use in several other applications, such as lecture segment labelling and summarisation.

8.1 Summary of Thesis Contributions

This section lists a summary of the contributions made by this thesis.

1. **Metadiscourse Tagging Approach for Academic Lectures:** A four-stage approach was developed to enrich academic lecture transcripts with metadiscourse, to improve the structure segmentation model of academic lectures. Most previous research has relied on the semantic-topical distribution of the discourse in segmenting the lecture, whereas the lecture's structure involves a mixture of topical and functional segments. For this reason, in this approach two different metadiscourse tagging models were developed, for purposes of comparison: a traditional classification model and a more advanced model based on a neural network. Both models were trained on a corpus annotated specifically for this task. Analysis of the results shows that the tagging task is domain-specific, as there is a difference in performance between Economics

and Physics lectures. The evaluation in Chapter 7 of the lecture structure segmentation task – using metadiscourse tags as features in the segmentation model, alongside lexical-cohesion features – shows that the presented approach achieves a significant improvement over the conventional models in segmenting lectures.

2. **A Corpus of Metadiscourse in Academic Lectures:** The first stage was introduced in Chapter 3, which delivered the metadiscourse corpus for academic lectures. Most previous works, particularly in the field of English language learning, do not contribute to corpora building when studying metadiscourse for lectures. Instead they provide a limited number of examples of the phenomenon, just to assist in the decision of categories selection. By contrast, studies in the field of natural language processing deliver several corpora on metadiscourse, each of which serves different tasks and domains other than lectures. For these reasons, there is a need to build a corpus of metadiscourse specifically for lecture courses, using an annotation scheme that fulfils the objective of this thesis of signalling different discourse functions. Experiments were completed with the help of expert annotators, to enable high-quality annotations; they achieved substantial inter-annotator agreement scores. The derived corpus is considered the core component in the metadiscourse tagging approach for the subsequent tasks.

3. **Automatic Transcriptions of Academic Lectures:** The second stage, described in Chapter 4, involved developing an Automatic Speech Recognition (ASR) system for academic lectures from different disciplines and on different topics, available online. The reference transcriptions used in previous annotation experiments were produced manually with the help of a transcribing service, which is a time-consuming task. In addition, they were often prone to transcription errors, such as when the transcriber was not aware of the technical words used and recorded other terminology instead. Any ASR system involves training acoustic and language models. However, due to mismatching between the training and test datasets, serious degradation is often observed in system performance. In this work, the developed ASR system focused on language model adaptation using linear interpolation of in-domain and out-of-domain resources. The ASR system of this stage was evaluated using the standard word error rate (WER). A lightly supervised alignment system is applied to the reference transcripts to provide time information, to facilitate the evaluation of the ASR outputs. Another benefit of this alignment system is the ability to correct some errors in the reference, since most of OCW platforms rely on commercial transcribing services that, as mentioned above, sometimes produce inaccurate transcriptions.

4. **Metadiscourse Tagging with SVMs:** The third stage of the tagging approach was the development of the first metadiscourse tagging model, in Chapter 5, using a

combination of textual and acoustic features. At this stage, the classification was accomplished using an SVM, due to its ability to combine high-dimensional (*e.g.,* word n-grams) with low-dimensional features (*e.g.,* prosodic cues). A number of test cases were set out to investigate the robustness of the tagging model, such as training and testing the model on both levels of granularities of metadiscourse tags: generic (only 4 tags) and specific (19 tags). In addition, to test the model on the ASR outputs that were produced in the previous stage, the strategy was followed to add the metadiscourse annotations from the reference transcription to the automatic ones. The model developed showed the ability to identify and classify metadiscourse instances in academic lectures. A downside of a model with such features representation is the sparsity problem, as the model was trained on a small annotated training dataset that does not capture all the variety of metadiscourse expressions. In order to function more effectively, a large labelled dataset is needed to cover all the different variants of metadiscourse expressions.

5. **Metadiscourse Tagging with CNNs:** The final stage of the tagging approach was an attempt to solve this issue by developing an alternative classification model, as it is not feasible to develop annotation for the large dataset required. This classification model utilises CNNs as a multiclass classification model, discussed in Chapter 6. At this stage, a continuous representation was used to represent both the textual and acoustic features. A key consideration in this model was how to obtain these continuous vectors, which require a large amount of data when training the network on them. For this reason, pre-trained word-embedding vectors from *word2vec* and *GloVe* were used to tune the network on these vectors, along with other features such as POS tags and prosodic cues. This model was evaluated using the same metric and experimental setup as in the previous model. Again, to prove the robustness of the model developed, a number of test cases were set out, including generic and specific metadiscourse tags. The results prove the robustness of the model, as it provides significantly better results than the SVM model. A further analysis of the results shows that this task is domain-specific, as was observed with the SVM model.

8.2 Directions for Future Research

The previous section describes our contributions towards the problem of metadiscourse tagging in academic lectures. There are some problems which still need to be addressed to improve the overall model performance. Future avenues for this work are discussed below.

Improving the Tagging Models

1. **Multi-domain Adaptation:** The metadiscourse tagging models in its current development relies on a small annotated dataset from both disciplines (Physics and Economics). The results obtained in this work indicate that metadiscourse tagging is a domain-specific task. For example, the classification results for Economics lectures were generally better than for Physics across a range of metadiscourse tags, which this study was able to show because the metadiscourse tagging experiment was carried out for each domain separately. However, developing a model for each new domain is not feasible, as this would require excessive annotation effort. In addition, using the training dataset from one domain to adapt for another domain is not possible because adaptation fails totally when the data distribution in the target domain is very different from that in the source domain. This problem could potentially be solved by utilising a multi-domain adaptation method that uses multiple domains in the training dataset of the source domain.

2. **Feature Vectors:** The proposed metadiscourse tagging model using CNNs has a solid performance in various test cases. However, it can still be improved. A key aspect of this model are the feature vectors used in training and testing for the classification task at hand. This work did not utilise the network to train these feature vectors, in order to produce continuous feature representations. Instead, pre-trained word embeddings were used from two sources: *word2vec* and *GloVe*. Other features types, such as POS tags, were handled by producing corpus-wide continuous values from the discrete representation, then tuning them for the task. The decision to follow this strategy in obtaining these continuous feature representations for multiple feature types was taken for practical reasons, in particular, because training the network to produce more tailored feature vectors for the task would require large amounts of labelled data. However, it may be worth investigating using an auxiliary task, such as a language model trained on huge amounts of unlabelled data, but having similar vocabulary (*e.g.,* textbooks, slides, lectures notes, *etc.*), to produce tuned feature vectors for the task. Other features, such as POS tags, can be treated in similar fashion to produce continuous feature vectors specificity for the classification task.

Other Application Areas:

1. **Summarisation:** So far, one way to evaluate the usefulness of metadiscourse tags with regards to downstream applications has been considered, which is lecture structure segmentation. Another application that could benefit from metadiscourse tags is summarisation, which is the process of reducing a lecture's content using a model to create a summary that contains most of the important concepts in the lecture. The idea

is for specific metadiscourse tags, such as *Emphasising* (EMP), to serve as indicators of important concepts in the discourse, which can be used as features in the summarisation model. When analysing the annotated corpus, it was noticed that EMP occurred quite frequently in both disciplines, with an average of 21.74 instances per lecture, as presented in Table 3.7. Clearly, enriching the transcripts with EMP tags will be beneficial for the summarisation task, in particular the extractive approach. Niekrasz (2012) shows that better understanding of the speech conversation type (*e.g.,* meeting) can be obtained by developing participant-relational features, which are a phenomenon that is similar to metadiscourse, and defined as expressions of relationships between participants and the dialogue. Their study argues that such features are important prerequisites to automatic summarisation of meeting based on its activities.

2. **Linking:** Another possible application regarding the use of metadiscourse is linking lecture courses that have very similar content. Such applications are becoming necessary with the increasing popularity of coherent sequences of lecture courses available online. The main objective in these applications is to give a new linked map of a good sequence order. For example, lecture courses often contain instances where lecturers link some concepts from the current lecture backwards or forwards to other lectures in the course. Tracking these instances and linking the concepts in the lecture course will not just organise these materials online, but also add to the learning process, as it provides a student with a map that allows them to understand the module effectively. To achieve this, one possible solution is to use the metadiscourse tags of *Previewing* (RRE) and *Reviewing* (REV) in lectures as indicators of these links in the course. For instance, based on the statistics of the annotated dataset, on average, Physics lecturers used 14.63 and 9.40 instances of REV or PRE per lecture, respectively. Similarly, in Economics lectures, they used on average 13.9 expression to review some concept and 8.08 to preview others, per lecture. Further, such linking application have been studied before, such as in Yang et al. (2015), but these studies measured the relation that forms the linking based on the similarity between lecture content, not its rhetorical functions. Using rhetorical functions in the form of metadiscourse tags to link online lecture courses has not yet been widely studied.

Appendix A

Annotation Instructions

Overview

In Economics lecture, lecturer often uses phrases to indicate what is important in lecture. These phrases are in the form of word or set of words which are not related to the topic of the lecture. In addition, these words are general and they often used to to indicate important or interesting. Examples of such phrases are: **"more important"**, **"what I want you to remember is"**, or **"I want to stress that"**. In this task, your job is to select word or set of words (similar to the highlighted phrases above) that lecturer use to indicate **important** in the provided lecture segments.

Rules and Tips

It is important to understand if the **highlighted** words indicate emphasis. **Do not** skim-reading the text you need to read carefully in order to inspect every sentence in the provided text. **Do not** select the whole sentence just the word or set of words that indicate emphasis/important/interesting.

Correct Examples

In the following you have set of correct examples of words that indicate **emphasis** and reasons for that:

Example1	**Now, what's very interesting is we've actually seen an evolution from my youth to your youth towards the economic model.**
Reason	Lecturer presents his interest about something in the lecture as shown in the highlighted phrase above. Your job is similar to this by highlighting the phrases that indicate important or interesting.
Example2	**The other distinction that's very important is positive versus normative economics.**
Reason	Here the lecturer clearly stats what is important using the highlighted phrase. Your job is similar to this by highlighting the phrases that indicate important or interesting.
Example3	**But the key point is you don't have to think about it exactly that way.**
Reason	The lecturer again shows clearly what is important using the highlighted phrase. Your job is similar to this by highlighting the phrases that indicate important or interesting.

Incorrect Examples

In the following you have set of incorrect examples of words that do not indicate **emphasis** and reasons for that:

Example1	**One other thing I want to cover before we stop which relates to the long run is that in the long run, this is sort of the transition, the bridge to talking about the long run.**
Reason	The lecturer here does not show neither important nor interest, he just use the highlighted phrase to inform what is coming next in the lecture.
Example2	**Distinction between the way things are and the way things should be. OK? So let's consider an example of microeconomics at work.**
Reason	The lecturer here uses the highlighted phrase to introduce example not to show important or interest.
Example3	**And what we'll learn this semester is about how you make them and how we model, how economists can use what we learn about that to understand the function of the economy.**
Reason	The lecturer here uses the highlighted phrase to preview what is coming next in the semester. He did not use phrase to indicate important or interesting.

Process

STEP1: Read the text below and click to mark/highlight words that you think is an indication of emphasis. Note that you can mark more than words in the text and you may also, not find any word to mark.

How to highlight words in the text:
- Click on each word in the text that you want to SELECT.
- Click on already highlighted word to DE-SELECT it.

STEP2: After Reading the text and mark words (if there is any). You need to confirm your decision by choosing one of the following options:

- The words that indicate emphasis in the text are now marked, or
- There is no occurrences in the text which indicate emphasis.

STEP3: Rate you confidence about the answer in a scale of 5, with 1 means that "you are not confident at all" and 5 "you are very confident".

FIGURE A.1: Example of the annotation instruction used in annotating the category *Emphasising*

Bibliography

Abbasi, A., H. Chen, and A. Salem (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Trans. Inf. Syst. 26*(3), 12:1–12:34.

Abdalla, R. M. and S. Teufel (2006). A bootstrapping approach to unsupervised detection of cue phrase variants. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, Stroudsburg, PA, USA, pp. 921–928. Association for Computational Linguistics.

Ädel, A. (2006). *Metadiscourse in L1 and L2 English*. Studies in corpus linguistics. John Benjamins Publishing Company.

Ädel, A. (2010). Just give kind of map of where we are going: A taxonomy of metadiscourse in spoken and written academic english. *Nordic Journal of English Studies 9*(2), 69–97.

Akita, Y., Y. Tong, and T. Kawahara (2015). Language model adaptation for academic lectures using character recognition result of presentation slides. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5431–5435. IEEE.

Alharbi, G. and T. Hain (2012). Automatic transcription of academic lectures from diverse disciplines. In *IEEE Spoken Language Technology Workshop (SLT)*, pp. 398–403.

Alharbi, G. and T. Hain (2015). Using topic segmentation models for the automatic organisation of moocs resources. In *The 8th International Conference on Education Data Mining (EDM)*, pp. 524–527.

Alharbi, G. and T. Hain (2016). The opencourseware metadiscourse (ocwmd) corpus. In *The 10th Edition of the Language Resources and Evaluation Conference (LREC)*.

Alharbi, G., R. W. M. Ng, and T. Hain (2015). Annotating meta-discourse in academic lectures from different disciplines. In *International Workshop on Speech and Language Technology in Education (SLaTE)*, pp. 161–166.

Asher, N. (2012). *Reference to abstract objects in discourse*, Volume 50. Springer Science & Business Media.

Auria, C. P.-L. (2006). Signaling speaker's intentions: towards a phraseology of textual metadiscourse in academic lecturing. *English as a GloCalization Phenomenon. Observations from a Linguistic Microcosm 3*, 59.

Bahl, L. R., F. Jelinek, and R. L. Mercer (1983). A maximum likelihood approach to continuous speech recognition. *IEEE transactions on pattern analysis and machine intelligence* (2), 179–190.

Bastien, F., P. Lamblin, R. Pascanu, J. Bergstra, I. J. Goodfellow, A. Bergeron, N. Bouchard, and Y. Bengio (2012). Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop.

Beeferman, D., A. Berger, and J. Lafferty (1999, February). Statistical models for text segmentation. *Mach. Learn. 34*(1-3), 177–210.

Bell, P., H. Yamamoto, P. Swietojanski, Y. Wu, F. McInnes, C. Hori, and S. Renals (2013). A lecture transcription system combining neural network acoustic and language models. In *The Annual Conference of the International Speech Communication Association (Interspeech)*, pp. 3087–3091.

Bellegarda, J. R. (2004). Statistical language model adaptation: review and perspectives. *Speech communication 42*(1), 93–108.

Bengio, Y., R. Ducharme, P. Vincent, and C. Janvin (2003, March). A neural probabilistic language model. *The Journal of Machine Learning Research 3*, 1137–1155.

Bergstra, J., O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio (2010). Theano: a cpu and gpu math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*.

Besling, S. and H.-G. Meier (1995). Language model speaker adaptation. In *Fourth European Conference on Speech Communication and Technology*.

Biber, D. (1986). Spoken and written textual dimensions in english: Resolving the contradictory findings. *Language*, 384–414.

Black, A. W. and N. Campbell (1995). Predicting the intonation of discourse segments from examples in dialogue speech. In *Spoken Dialogue Systems-Theories and Applications*.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003, March). Latent dirichlet allocation. *J. Mach. Learn. Res. 3*, 993–1022.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *IFA Proceedings 17*, pp. 97–110.

Bulyko, I., M. Ostendorf, and A. Stolcke (2003). Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Companion Volume of the Proceedings of HLT-NAACL 2003–short Papers - Volume 2*, NAACL-Short '03, pp. 7–9.

Burger, J., D. Palmer, and L. Hirschman (1998, August). Named entity scoring for speech input. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, Montreal, Quebec, Canada, pp. 201–205. Association for Computational Linguistics.

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics 22*(2).

Carlson, L. and D. Marcu (2001). Discourse tagging reference manual. *ISI Technical Report ISI-TR-545*.

Carlson, L., D. Marcu, and M. E. Okurowski (2003). Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Current and new directions in discourse and dialogue*, pp. 85–112. Springer.

Catarina, S. and R. Bernardete (2003). The importance of stop word removal on recall values in text categorization. In *Proceedings of the International Joint Conference on Neural Networks*, Volume 3, pp. 1661–1666.

Cerva, P., J. Silovský, J. Zdánský, J. Nouza, and J. Málek (2012). Real-time lecture transcription using asr for czech hearing impaired or deaf students. In *The Annual Conference of the International Speech Communication Association (Interspeech)*, pp. 763–766. ISCA.

Choi, F. Y. Y. (2000). Advances in domain independent linear text segmentation. In *Proceedings of NAACL'00*, pp. 26–33.

Clarkson, P. R. and A. J. Robinson (1997). Language model adaptation using mixtures and an exponentially decaying cache. In *The IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Volume 2, pp. 799–802. IEEE.

Cohen, W. W., V. R. Carvalho, and T. M. Mitchell (2004). Learning to classify email into "speech acts". In D. Lin and D. Wu (Eds.), *Proceedings of EMNLP 2004*, Barcelona, Spain, pp. 309–316. Association for Computational Linguistics.

Collobert, R. and J. Weston (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pp. 160–167. ACM.

Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa (2011). Natural language processing (almost) from scratch. *The Journal of Machine Learning Research 12*, 2493–2537.

Core, M. G. and J. F. Allen (1997). Coding dialogs with the damsl annotation scheme.

Correia, R., M. Eskenazi, and N. Mamede (2015, September). Lexical level distribution of metadiscourse in spoken language. In *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*, Lisbon, Portugal, pp. 70–75. Association for Computational Linguistics.

Correia, R., N. Mamede, J. Baptista, and M. Eskenazi (2014a). Toward automatic classification of metadiscourse. In *Advances in Natural Language Processing*, pp. 262–269. Springer International Publishing.

Correia, R., N. Mamede, J. Baptista, and M. Eskenazi (2014b). Using the crowd to annotate metadiscursive acts. *In Proceedings 10th Joint ISO-ACL SIGSEM*, 102.

Correia, R., N. Mamede, J. Baptista, and M. Eskenazi (2016). metated: a corpus of metadiscourse for spoken language. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Correia, R. P. d. S. (2013). *Automatic Classification of Metadiscourse for Presentation Skills Instruction*. Ph. D. thesis, Thesis Proposal University of Carnegie Mellon.

Cortes, C. and V. Vapnik (1995, September). Support-vector networks. *Machine Learning 20*(3), 273–297.

Crismore, A. (1989). Talking with readers: Metadiscourse as rhetorical act. *17*.

Crismore, A., R. Markkanen, and M. S. Steffensen. (1993). Metadiscourse in persuasive writing a study of texts written by American and Finnish university students. *Written communication 10*(2), 39–71.

Dahl, G. E., D. Yu, L. Deng, and A. Acero (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing 20*(1), 30–42.

Davis, J. and M. Goadrich (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240. ACM.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1–38.

Du, L., W. L. Buntine, and M. Johnson (2013). Topic segmentation with a structured topic model. In *HLT-NAACL*, pp. 190–200.

Eisenstein, J. and R. Barzilay (2008). Bayesian unsupervised topic segmentation. *Proceedings of EMNLP'08*, 334.

Elman, J. L. (1990). Finding structure in time. *COGNITIVE SCIENCE 14*(2), 179–211.

Entropic (1993). Esps version 5.0 programs manual.

Erhan, D., P.-A. Manzagol, Y. Bengio, S. Bengio, and P. Vincent (2009). The difficulty of training deep architectures and the effect of unsupervised pre-training. In *International Conference on artificial intelligence and statistics*, pp. 153–160.

Fan, R.-E., K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin (2008). Liblinear: A library for large linear classification. *The Journal of Machine Learning Research 9*, 1871–1874.

Federico, M., M. Cettolo, L. Bentivogli, M. Paul, and S. Stüker (2012). Overview of the iwslt 2012 evaluation campaign. In *International Workshop on Spoken Language Translation (IWSLT)*.

Fernandez, R. and R. W. Picard (2002). Dialog act classification from prosodic features using support vector machines. In *International Conference on Speech Prosody*.

Finke, M., M. Lapata, A. Lavie, L. Levin, L. M. Tomokiyo, T. Polzin, K. Ries, A. Waibel, and K. Zechner (1998). Clarity: Inferring discourse structure from speech. In *In Proceedings of Workshop on Applying Machine Learning to Discourse Processing*.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin 76*(5), 378.

Furui, S. (2003). Recent advances in spontaneous speech recognition and understanding. In *ISCA & IEEE workshop on spontaneous speech processing and recognition*.

Furui, S., K. Maekawa, H. Isahara, T. Shinozaki, and T. Ohdaira (2000). Toward the realization of spontaneous speech recognition-introduction of a japanese priority program and preliminary results. In *The Sixth International Conference on Spoken Language Processing*, pp. 518–521.

Galibert, O., S. Rosset, C. Grouin, P. Zweigenbaum, and L. Quintard (2011, November). Structured and extended named entity evaluation in automatic speech transcriptions. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, Chiang Mai, Thailand, pp. 518–526. Asian Federation of Natural Language Processing.

Galley, M., K. McKeown, E. Fosler-Lussier, and H. Jing (2003). Discourse segmentation of multi-party conversation. In *Proceedings of ACL'03*, pp. 562–569.

Georgescul, M., A. Clark, and S. Armstrong (2006). Word distributions for thematic segmentation in a support vector machine approach. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pp. 101–108. Association for Computational Linguistics.

Georgescul, M., A. Clark, and S. Armstrong (2007). Exploiting structural meeting-specific features for topic segmentation.

Gibson, M. and T. Hain (2006). Hypothesis spaces for minimum bayes risk training in large vocabulary speech recognition. In *The Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 2–4.

Glass, J. (2003). A probabilistic framework for segment-based speech recognition. *Computer Speech and Language 17*(2), 137–152.

Glass, J., T. J. Hazen, S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay (2007). Recent progress in the mit spoken lecture processing project. In *The Eighth Annual Conference of the International Speech Communication Association (Interspeech)*, pp. 2553–2556.

Glass, J., T. J. Hazen, L. Hetherington, and C. Wang (2004). Analysis and processing of lecture audio data: Preliminary investigations. In *Proceedings of the Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval at HLT-NAACL 2004*, pp. 9–12. Association for Computational Linguistics.

Goldberg, Y. (2015). A primer on neural network models for natural language processing. *arXiv preprint arXiv:1510.00726*.

Goodman, J. T. (2001). A bit of progress in language modeling. *Computer Speech & Language 15*(4), 403–434.

Grosz, B. J. and C. L. Sidner (1986, July). Attention, intentions, and the structure of discourse. *Comput. Linguist. 12*(3), 175–204.

Hain, T., L. Burget, J. Dines, P. N. Garner, F. Grézl, A. E. Hannani, M. Huijbregts, M. Karafiat, M. Lincoln, and V. Wan (2012). Transcribing meetings with the amida systems. *Audio, Speech, and Language Processing, IEEE Transactions on 20*(2), 486–498.

Hain, T., J. Dines, G. Garau, M. Karafiát, D. Moore, V. Wan, R. Ordelman, and S. Renals (2005). Transcription of conference room meetings: an investigation. In *The 9th European Conference on Speech Communication and Technology (INTERSPEECH), Lisbon, Portugal*, pp. 1661–1664.

Hasler, E., P. Bell, A. Ghoshal, B. Haddow, P. Koehn, F. McInnes, S. Renals, and P. Swietojanski (2012). The uedin systems for the iwslt 2012 evaluation. In *International Workshop on Spoken Language Translation (IWSLT)*.

Hayes, P. J., A. G. Hauptmann, G. Carbonell, Jaime, and M. Tomita (1986). Parsing spoken language: A semantic caseframe approach. In *Proceedings of the 11th International Conference on Computational Linguistics, COLING'86, Bonn, Germany, August 25-29, 1986*, pp. 587–592.

Hazen, T. J. (2006). Automatic alignment and error correction of human generated transcripts for long speech recordings.

Hearst, M. A. (1997, March). Texttiling: Segmenting text into multi-paragraph subtopic passages. *Comput. Linguist. 23*(1), 33–64.

Hendrickx, I., S. N. Kim, Z. Kozareva, P. Nakov, D. Ó Séaghdha, S. Padó, M. Pennacchiotti, L. Romano, and S. Szpakowicz (2009). Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pp. 94–99. Association for Computational Linguistics.

Hinton, G., L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE 29*(6), 82–97.

Hinton, G. E., N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov (2012). Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*.

Hirschberg, J. and D. Litman (1993a). Empirical studies on the disambiguation of cue phrases. *Computational linguistics 19*(3), 501–530.

Hirschberg, J. and D. Litman (1993b). Empirical studies on the disambiguation of cue phrases. *Comput. Linguist. 19*(3), 501–530.

Hirschman, L., J. Burger, D. Palmer, and P. Robinson (1999). Evaluating content extraction from audio sources. In *in ECSA, ETRW Workshop: Accessing Infomation in Spoken Audio*. University Press.

Hockett, C. F. (1963). *The problem of universals in language*, Volume 2. Universals of language.

Hsu, B.-J. P. and J. Glass (2006, July). Style & topic language model adaptation using hmm-lda. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia, pp. 373–381. Association for Computational Linguistics.

Hsueh, P.-Y., J. D. Moore, and S. Renals (2006). Automatic segmentation of multiparty dialogue. In *Proceedings of the 11th Conference of EACL*.

Hu, B., Z. Lu, H. Li, and Q. Chen (2014). Convolutional neural network architectures for matching natural language sentences. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 27*, pp. 2042–2050. Curran Associates, Inc.

Hu, M. and B. Liu (2004). Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, New York, NY, USA, pp. 168–177. ACM.

Huang, F. J., Y.-L. Boureau, Y. LeCun, et al. (2007). Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *2007 IEEE conference on computer vision and pattern recognition*, pp. 1–8. IEEE.

Hubel, D. and T. Wiesel (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology 195*(1), 215–243.

Hyland, K. (1998). Persuasion and context: The pragmatics of academic metadiscourse. *Journal of pragmatics 30*(4), 437–455.

Hyland, K. (2005). *Metadiscourse: Exploring Interaction in Writing*. Continuum Discourse. Bloomsbury Publishing.

Iyer, R. M. and M. Ostendorf (1999). Modeling long distance dependence in language: Topic mixtures versus dynamic cache models. *IEEE Transactions on speech and audio processing 7*(1), 30–39.

Jelinek, F., R. L. Mercer, L. R. Bahl, and J. K. Baker (1977). Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America 62*(S1), S63–S63.

Joachims, T. (1998). Text categorization with suport vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning*, ECML '98, London, UK, pp. 137–142. Springer-Verlag.

Jones, D., F. Wolf, E. Gibson, E. Williams, E. Fedorenko, D. Reynolds, and M. Zissman (2003). Measuring the readability of automatic speech-to-text transcripts. In *The 8th European Conference on Speech Communication and Technology, EUROSPEECH INTERSPEECH, Geneva, Switzerland, September*.

Jones, D.and Gibson, E., W. Shen, N. Granoien, M. Herzog, D. Reynolds, and C. Weinstein (2005). Measuring human readability of machine generated text: three case studies in speech recognition and machine translation. In *The IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP)*, Volume 5.

Kalchbrenner, N., E. Grefenstette, and P. Blunsom (2014, June). A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Baltimore, Maryland, pp. 655–665. Association for Computational Linguistics.

Katsamanis, A., M. P. Black, P. Georgiou, L. Goldstein, and S. S. Narayanan (2011). Sailalign: Robust long speech-text alignment. In *Proc. of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*.

Kauchak, D. and F. Chen (2005). Feature-based segmentation of narrative documents. In *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, pp. 32–39. Association for Computational Linguistics.

Kim, J.-K., M.-C. de Marneffe, and E. Fosler-Lussier (2015). Neural word embeddings with multiplicative feature interactions for tensor-based compositions. In *Proceedings of NAACL-HLT*, pp. 143–150.

Kim, S.-M., P. Pantel, T. Chklovski, and M. Pennacchiotti (2006). Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pp. 423–430.

Kim, Y. (2014, October). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, pp. 1746–1751. Association for Computational Linguistics.

Kingsbury, B., T. N. Sainath, and H. Soltau (2012). Scalable minimum bayes risk training of deep neural network acoustic models using distributed hessian-free optimization. In *The 13th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 10–13.

Klakow, D. and J. Peters (2002). Testing the correlation of word error rate and perplexity. *Speech Communication 38*(1), 19–28.

Klein, D. and C. D. Manning (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, Stroudsburg, PA, USA, pp. 423–430. Association for Computational Linguistics.

Kokhlikyan, N., A. Waibel, Y. Zhang, and J. Y. Zhang (2013). Measuring the structural importance through rhetorical structure index. *NAACL HLT 2013*, 783–788.

Kombrink, S., T. Mikolov, M. Karafiát, and L. Burget (2012). Improving language models for asr using translated in-domain data. In *The IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4405–4408. IEEE.

Kompe, R., J. Siekmann, and J. Carbonell (1997). *Prosody in speech understanding systems.* Springer-Verlag New York, Inc.

Kopple, W. J. V. (1985). Some exploratory discourse on metadiscourse. *College composition and communication*, 82–93.

Kuhn, R. and R. De Mori (1990). A cache-based natural language model for speech recognition. *IEEE transactions on pattern analysis and machine intelligence 12*(6), 570–583.

Lamel, L., J.-L. Gauvain, and G. Adda (2002). Lightly supervised and unsupervised acoustic model training. *Computer Speech and Language 16*(1), 115–129.

Lamel, L. F., F. Schiel, A. Fourcin, J. Mariani, and H. G. Tillmann (1994). The translanguage english database (ted). In *Third International Conference on Spoken Language Processing.*

Landis, J. R. and G. G. Koch (1977). The measurement of observer agreement for categorical data. *biometrics*, 159–174.

Lecun, Y. and Y. Bengio (1995). *Convolutional networks for images, speech, and time-series.* MIT Press.

LeCun, Y., L. Bottou, Y. Bengio, and P. Haffner (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE 86*(11), 2278–2324.

LeCun, Y., S. Chopra, and R. Hadsell (2006). A tutorial on energy-based learning.

Lee, Y. K. and H. T. Ng (2002). An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA, pp. 41–48. Association for Computational Linguistics.

Levy, R. and G. Andrew (2006). Tregex and tsurgeon: tools for querying and manipulating tree data structures. In *The fifth international conference on Language Resources and Evaluation 2006*, pp. 2231–2234.

Lewis, D. D., Y. Yang, T. G. Rose, and F. Li (2004, December). Rcv1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res. 5*, 361–397.

Li, X. and D. Roth (2002). Learning question classifiers. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, COLING '02, Stroudsburg, PA, USA, pp. 1–7. Association for Computational Linguistics.

Lioma, C. and I. Ounis (2006). Examining the content load of part of speech blocks for information retrieval. In *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, COLING-ACL '06, Stroudsburg, PA, USA, pp. 531–538. Association for Computational Linguistics.

Litman, D. J. and R. J. Passonneau (1995). Combining multiple knowledge sources for discourse segmentation. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pp. 108–115. Association for Computational Linguistics.

Liu, Y., E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper (2006). Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *Audio, Speech, and Language Processing, IEEE Transactions on 14*(5), 1526–1540.

Lucy, J. (1993). *Reflexive Language: Reported Speech and Metapragmatics*. Cambridge University Press.

Luukka, M.-R. (1992). Metadiscourse in academic texts. In *Conference on Discourse and the Professions*, Volume 28, Uppsala, Sweden.

Lyons, J. (1977). *Semantics::*. Semantics. Cambridge University Press.

Maas, A. L., R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150. Association for Computational Linguistics.

Madnani, N., M. Heilman, J. Tetreault, and M. Chodorow (2012). Identifying high-level organizational elements in argumentative discourse. In *Proceedings of NAACL'12: HLT*, pp. 20–28.

Malioutov, I. and R. Barzilay (2006). Minimum cut model for spoken lecture segmentation. In *Proceedings ACL '06*, pp. 25–32.

Mann, W. C. and S. A. Thompson (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text 8*(3), 243–281.

Marcu, D. (2000). *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press.

Marcu, D. and A. Echihabi (2002). An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 368–375. Association for Computational Linguistics.

Marcus, M. P., M. A. Marcinkiewicz, and B. Santorini (1993). Building a large annotated corpus of english: The penn treebank. *Computational linguistics 19*(2), 313–330.

Martínez-Villaronga, A., M. del Agua, J. A. Silvestre-Cerdà, J. Andrés-Ferrer, and A. Juan (2014). Language model adaptation for lecture transcription by document retrieval. In *Advances in Speech and Language Technologies for Iberian Languages*, pp. 129–137. Springer.

Martínez-Villaronga, A., A. Miguel, J. Andrés-Ferrer, and A. Juan (2013). Language model adaptation for video lectures transcription. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8450–8454. IEEE.

Mauranen, A. (1993). Cultural differences in academic discourse–problems of a linguistic and cultural minority. *AFinLA Yearbook 23*(51), 157–174.

Mauranen, A. (2001). Reflexive academic talk: Observations from micase. In *Corpus linguistics in North America Selections from the 1999 symposium*, pp. 165–178.

McLachlan, G. and T. Krishnan (2007). *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics. Wiley.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119.

Milajevs, D., D. Kartsaklis, M. Sadrzadeh, and M. Purver (2014). Evaluating neural word representations in tensor-based compositional settings. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar. Association for Computational Linguistics.

Miltsakaki, E., L. Robaldo, A. Lee, and A. Joshi (2008). Sense annotation in the penn discourse treebank. In *Proceedings of the LREC'08*.

Misra, H., F. Yvon, J. M. Jose, and O. Cappe (2009). Text segmentation via topic modeling: An analytical study. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pp. 1553–1556.

Mohri, M. (2004). Weighted finite-state transducer algorithms. an overview. In *Formal Languages and Applications*, pp. 551–563. Springer.

Mohri, M., P. Moreno, and E. Weinstein (2010). Discriminative topic segmentation of text and speech. In *International Conference on Artificial Intelligence and Statistics*, pp. 533–540.

Mohri, M., F. Pereira, and M. Riley (2002). Weighted finite-state transducers in speech recognition. *Computer Speech and Language 16*(1), 69–88.

Moniz, H., F. Batista, I. Trancoso, and A. I. Mata (2012). Prosodic context-based analysis of disfluencies. In *The Annual Conference of the International Speech Communication Association (Interspeech)*, Portland, Oregon, USA. ISCA.

Moore, R. C. and W. Lewis (2010). Intelligent selection of language model training data. In *Proceedings of the ACL 2010 conference short papers*, pp. 220–224. Association for Computational Linguistics.

Morris, J. and G. Hirst (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational linguistics 17*(1), 21–48.

Mou, L., H. Peng, G. Li, Y. Xu, L. Zhang, and Z. Jin (2015, September). Discriminative neural sentence modeling by tree-based convolution. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, pp. 2315–2325. Association for Computational Linguistics.

Mullen, T. and N. Collier (2004). Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain, pp. 412–418. Association for Computational Linguistics.

Munteanu, C., R. Baecker, and G. Penn (2008). Collaborative editing for improved usefulness and usability of transcript-enhanced webcasts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 373–382. ACM.

Nair, V. and G. E. Hinton (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 807–814.

Niekrasz, J. (2012). *Toward Summarization of Communicative Activities in Spoken Conversation*. Ph. D. thesis, University of Edinburgh.

NIST (2009). National institute of standards and technology toolkit. https://www.nist.gov/itl/iad/mig/tools. [Online; accessed 19-January-2016].

Novak, J. R., N. Minematsu, and K. Hirose (2012). Wfst-based grapheme-to-phoneme conversion: Open source tools for alignment, model-building and decoding. In *10th International Workshop on Finite State Methods and Natural Language Processing*, pp. 45. Citeseer.

Olcoz, J., O. Saz, and T. Hain (2016). Error correction in lightly supervised alignment of broadcast subtitles. In *Proceedings of the 17th Annual Conference of the International Speech Communication Association (Interspeech)*, San Francisco, CA.

Osińsk, S. and D. Weiss (2005). A concept-driven algorithm for clustering search results. *IEEE Intelligent Systems 20*(3), 48–54.

Osiński, S., J. Stefanowski, and D. Weiss (2004). Lingo: Search results clustering algorithm based on singular value decomposition. pp. 359–368. Springer.

Paltoglou, G. and M. Thelwall (2010). A study of information retrieval weighting schemes for sentiment analysis. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA, pp. 1386–1395. Association for Computational Linguistics.

Pang, B. and L. Lee (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, pp. 271. Association for Computational Linguistics.

Pang, B. and L. Lee (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, Stroudsburg, PA, USA, pp. 115–124. Association for Computational Linguistics.

Pang, B., L. Lee, and S. Vaithyanathan (2002). Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, Stroudsburg, PA, USA, pp. 79–86. Association for Computational Linguistics.

Passonneau, R. J. and D. J. Litman (1997). Discourse segmentation by human and automated means. *Computational Linguistics 23*(1), 103–139.

Pedersen, T. (2001). A decision tree of bigrams is an accurate predictor of word sense. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, NAACL '01, Stroudsburg, PA, USA, pp. 1–8. Association for Computational Linguistics.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12*, 2825–2830.

Pennington, J., R. Socher, and C. Manning (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, pp. 1532–1543. Association for Computational Linguistics.

Pevzner, L. and M. A. Hearst (2002, March). A critique and improvement of an evaluation metric for text segmentation. *Comput. Linguist. 28*(1), 19–36.

Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, pp. 61–74. MIT Press.

Povey, D., A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, Number EPFL-CONF-192584. IEEE Signal Processing Society.

Povey, D., D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah (2008). Boosted mmi for model and feature-space discriminative training. In *The IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4057–4060. IEEE.

Prasad, R., N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber (2008). The penn discourse treebank 2.0. In *In Proceedings of LREC*.

Punyakanok, V., D. Roth, and W.-t. Yih (2008). The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics 34*(2), 257–287.

Purver, M. (2011). Topic segmentation. *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, 291–317.

Qadir, A. and E. Riloff (2011, July). Classifying sentences as speech acts in message board posts. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK., pp. 748–758. Association for Computational Linguistics.

Ramamurti, S. (2006). Fundamental of physics 1 (yale university: Open yale courses). http://oyc.yale.edu/physics/phys-200#overview. (Accessed December 20, 2014), License: Creative Commons BY-NC-SA.

Riedl, M. and C. Biemann (2012). Topictiling: A text segmentation algorithm based on lda. In *Proceedings of ACL'12 Student Research Workshop*, pp. 37–42.

Rifkin, R. and A. Klautau (2004). In defense of one-vs-all classification. *The Journal of Machine Learning Research 5*, 101–141.

Rijsbergen, C. J. V. (1979). *Information Retrieval* (2nd ed.). Butterworth-Heinemann.

Riloff, E. (1993). Automatically constructing a dictionary for information extraction tasks. In *Proceedings of the Eleventh National Conference on Artificial Intelligence*, AAAI'93, pp. 811–816. AAAI Press.

Roṁer, U. and J. M. Swales (2009). The michigan corpus of upper-level student papers (micusp). *Journal of English for Academic Purposes*.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams (1988). Learning representations by back-propagating errors. *Cognitive modeling 5*(3), 1.

Salton, G. and M. J. McGill (1986). *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, Inc.

Salton, G., A. Wong, and C. S. Yang (1975, November). A vector space model for automatic indexing. *Communications of the ACM 18*(11), 613–620.

Saz, O., M. Doulaty, S. Deena, R. Milner, R. Ng, M. Hasan, Y. Liu, and T. Hain (2015). The 2015 sheffield system for transcription of multi–genre broadcast media. In *Proceedings of the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, AZ.

Scaiano, M. and D. Inkpen (2012). Getting more from segmentation evaluation. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pp. 362–366. Association for Computational Linguistics.

Schiffrin, D. (1980). Meta-talk: Organizational and evaluative brackets in discourse. *Socioogical Inquiry 50*(3–4), 199–236.

Schmid, H. (1994). Part-of-speech tagging with neural networks. In *Proceedings of the 15th Conference on Computational Linguistics*, pp. 172–176. Association for Computational Linguistics.

Seide, F., G. Li, X. Chen, and D. Yu (2011). Feature engineering in context-dependent deep neural networks for conversational speech transcription. In *Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 24–29. IEEE.

Shen, Y., X. He, J. Gao, L. Deng, and G. Mesnil (2014). Learning semantic representations using convolutional neural networks for web search. In *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, pp. 373–374. International World Wide Web Conferences Steering Committee.

Shiller, R. J. (2011). Financial markets (yale university: Open yale courses). http://oyc.yale.edu/economics/econ-252-11. (Accessed December 22, 2014), License: Creative Commons BY-NC-SA.

Shriberg, E., A. Stolcke, D. Hakkani-Tür, and G. Tür (2000, September). Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication 32*(1-2), 127–154.

Shriberg, E., A. Stolcke, D. Jurafsky, N. Coccaro, M. Meteer, R. Bates, P. Taylor, K. Ries, R. Martin, and C. Van Ess-Dykema (1998). Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and speech 41*(3-4), 443–492.

Simpson, Rita C., B. S. L. O. J. and J. M. Swales (2002). The michigan corpus of academic spoken english.

Socher, R., A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, Volume 1631, pp. 1642. Citeseer.

Stan, A., Y. Mamiya, J. Yamagishi, P. Bell, O. Watts, R. Clark, and S. King (2016). ALISA: An automatic lightly supervised speech segmentation and alignment tool. *Computer Speech and Language 35*, 116–133.

Steinkrau, D., P. Y. Simard, and I. Buck (2005). Using gpus for machine learning algorithms. In *The 12th International Conference on Document Analysis and Recognition (ICDAR 2005)*, pp. 1115–1119.

Stolcke, A. (2002). Srilm-an extensible language modeling toolkit. In *The 7th International Conference on Spoken Language Processing (ICSLP)*, pp. 901–904.

Stolcke, A., N. Coccaro, R. Bates, P. Taylor, C. Van Ess-Dykema, K. Ries, E. Shriberg, D. Jurafsky, R. Martin, and M. Meteer (2000, September). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics 26*(3), 339–373.

Stouten, F., J. Duchateau, J.-P. Martens, and P. Wambacq (2006, 11). Coping with disfluencies in spontaneous speech recognition: acoustic detection and linguistic context manipulation. *Speech Communication 48*, 1590–1606.

Sun, Q., R. Li, D. Luo, and X. Wu (2008). Text segmentation with lda-based fisher kernel. In *Proceedings of ACL'08: Short Papers*, pp. 269–272.

Sundermeyer, M., Z. Tüske, R. Schlüter, and H. Ney (2014). Lattice decoding and rescoring with long-span neural network language models. In *The Annual Conference of the International Speech Communication Association (Interspeech)*, pp. 661–665.

Surendran, D. and G.-a. Levow (2006). Dialog act tagging with support vector machines and hidden markov models. In *The Annual Conference of the International Speech Communication Association (Interspeech)*.

Teufel, S. (1998). Meta-discourse markers and problem-structuring in scientific articles. In *Proceedings of the Workshop on Discourse Relations and Discourse Markers at the 17th International Conference on Computational Linguistics*, pp. 43–49.

Teufel, S. and M. Moens (2002). Summarizing scientific articles - experiments with relevance and rhetorical status. *Computational Linguistics 28*, 2002.

Thompson, S. E. (2003). Text-structuring metadiscourse, intonation and the signalling of organisation in academic lectures. *Journal of English for academic purposes 2*(1), 5–20.

Tsuboi, Y. (2014, October). Neural networks leverage corpus-wide information for part-of-speech tagging. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, pp. 938–950. Association for Computational Linguistics.

Tür, G., D. Hakkani-Tür, A. Stolcke, and E. Shriberg (2001). Integrating prosodic and lexical cues for automatic topic segmentation. *Computational linguistics 27*(1), 31–57.

Turney, P. D. and P. Pantel (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research 37*(1), 141–188.

Utiyama, M. and H. Isahara (2001). A statistical model for domain-independent text segmentation. In *Proceedings of ACL'01*, pp. 499–506.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag New York, Inc.

Vasile, R., M. Philip M., L. Mihai C., M. Danielle S., and G. Arthur C. (2008). Paraphrase identification with lexico-syntactic graph subsumption. In *Proceedings of the Twenty-First International Florida Artificial Intelligence Research Society Conference, May 15-17, 2008, Coconut Grove, Florida, USA*, pp. 201–206.

Venkataraman, A., A. Stolcke, and E. Shriberg (2002). Automatic dialog sct labeling with minimal supervision. In *Proceedings of the 9th Australian International Conference on Speech Science and Technology*, Melbourne, Australia.

Veselỳ, K., L. Burget, and F. Grézl (2010). Parallel training of neural networks for speech recognition. In *Text, Speech and Dialogue*, pp. 439–446. Springer.

W. M. Ng, R., M. Doulaty, R. Doddipatla, W. Aziz, K. Shah, O. Saz, M. Hasan, G. AlHarbi, L. Specia, and T. Hain (2015). The USFD spoken language translation system for IWSLT 2014.

Wang, X. and A. McCallum (2006). Topics over time: A non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, pp. 424–433. ACM.

Webb, N., M. Hepple, and Y. Wilks (2005). Dialogue act classification based on intra-utterance features. In *Proceedings of the AAAI Workshop on Spoken Language Understanding 2005*. Citeseer.

Webber, B. (2004). D-ltag: extending lexicalized tag to discourse. *Cognitive Science 28*(5), 751–779.

Webber, B., A. Joshi, E. Miltsakaki, R. Prasad, N. Dinesh, A. Lee, and K. Forbes (2005). A short introduction to the penn discourse treebank. In *In Copenhagen Working Papers in Language and Speech Processing*. Citeseer.

Wiebe, J., T. Wilson, and C. Cardie (2005). Annotating expressions of opinions and emotions in language. *Language resources and evaluation 39*(2-3), 165–210.

Wilks, Y. and M. Stevenson (1998). The grammar of sense: Using part-of-speech tags as a first step in semantic disambiguation. *Natural Language Engineering 4*(2), 135–143.

Wilson, S. (2010). Distinguishing use and mention in natural language. In *Proceedings of the NAACL HLT 2010 Student Research Workshop*, pp. 29–33. Association for Computational Linguistics.

Wilson, S. (2012). The creation of a corpus of english metalanguage. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pp. 638–646. Association for Computational Linguistics.

Wilson, S. (2013, October). Toward automatic processing of english metalanguage. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, Nagoya, Japan, pp. 760–766. Asian Federation of Natural Language Processing.

Wilson, T., J. Wiebe, and P. Hoffmann (2009). Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational linguistics 35*(3), 399–433.

Xiaodan, Z. and P. Gerald (2006). Summarization of spontaneous conversations. In *in Proceedings of the Annual Conference of the International Speech Communication Association*, pp. 1531–1534.

Xiong, W. and D. Litman (2010). Identifying problem localization in peer-review feedback. In *Intelligent Tutoring Systems*, pp. 429–431. Springer.

Xu, J. and W. B. Croft (1998). Corpus-based stemming using cooccurrence of word variants. *ACM Transactions on Information Systems (TOIS) 16*(1), 61–81.

Yamamoto, H., Y. Wu, C.-L. Huang, X. Lu, P. R. Dixon, S. Matsuda, C. Hori, and H. Kashioka (2012). The nict asr system for iwslt2012. In *International Workshop on Spoken Language Translation (IWSLT) 2012*.

Yang, Y., H. Liu, J. Carbonell, and W. Ma (2015). Concept graph learning from educational data. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pp. 159–168. ACM.

Yih, W.-t., X. He, and C. Meek (2014, June). Semantic parsing for single-relation question answering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Baltimore, Maryland, pp. 643–648. Association for Computational Linguistics.

Young, S., G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, et al. (2006). *The HTK book*.

Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Zeng, D., K. Liu, S. Lai, G. Zhou, and J. Zhao (2014a, August). Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, Dublin, Ireland, pp. 2335–2344. Dublin City University and Association for Computational Linguistics.

Zeng, D., K. Liu, S. Lai, G. Zhou, and J. Zhao (2014b, August). Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 2335–2344.

Zhang, Y. and B. Wallace (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *CoRR*.

Zimmerman, D. W. (1997). A note on interpretation of the paired-samples t test. *Journal of Educational and Behavioral Statistics*, 349–360.